



Abioye, Jumai Adeola (2018) *Engineering chimaeric recombinases for HIV-1 proviral DNA excision*. PhD thesis.

<https://theses.gla.ac.uk/9143/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

Engineering Chimaeric Recombinases for HIV-1 Proviral DNA Excision

Jumai Adeola Abioye
B.Sc. (Hons.), M.Sc.

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

May 2018

©J. A. Abioye, 2018

This thesis is lovingly dedicated to my parents.
I am not the type of doctor you originally hoped for,
but here I am, a Dr still.

I am only one, but I am one.
I cannot do everything, but I can do something.
The something I ought to do, I can do.
And by the grace of God, I will.

- Edward Everett Hale

Abstract

'Cutting out' HIV-1 proviral DNA could potentially cure a person of the infection. Genome editing approaches have been proffered for eradicating the provirus in infected persons by activating all latent viral reservoirs for further antiretroviral therapy or for the excision of the proviral DNA from memory T- cells. Previous approaches to do this have used nuclease-based tools or reprogrammed tyrosine recombinases; the former presenting unpredictable therapeutic outcomes and the latter, lengthy design time for newer tool variants if viral mutability erodes their effectiveness. Unlike nuclease-based tools that only cut DNA and rely on host-mediated repair mechanisms, chimaeric recombinases (CRs) cut DNA and carry the inherent ability to re-ligate cut ends at the cleavage site. The modular domain architecture of small serine recombinases can be redesigned to mediate site-specific recombination on non-cognate sites, by replacing the C-terminal DNA binding domains (DBDs) of serine recombinases with programmable DBDs such as Zinc Finger (ZF) proteins, TAL effector proteins and CRISPR-dCas9. For HIV-1 proviral DNA excision, CR requirement for interaction of two recombinase-bound sites, and the lack of necessity for host cell-encoded factors should maximize the fidelity and efficiency of provirus removal.

In this work, the engineering and characterization of CRs with the specificity to recognize and promote site-specific recombination at a highly conserved region within the HIV-1 long terminal repeats (LTR) flanking the proviral DNA at both ends is explored. First, a combination of random mutagenesis and rational engineering approaches was applied to create Tn3 resolvase-based recombinase catalytic domain variants that catalyse efficient excision at the chosen HIV-1 target sequence in a mock-HIV substrate construct in *Escherichia coli* cells. These Zif268-guided recombinase catalytic domain mutants were directed to HIV Z-sites that consist of a 16-bp HIV LTR sequence flanked by Zif268 target sequences. To generate completely reprogrammed HIV-excising TAL effector CRs (TALERS), the architecture and design of previously-described functional TALERS needed to be optimised. This was achieved by testing various fusion constructs of truncated TAL effector DBDs with the Tn3 resolvase catalytic domain, analysing the linker lengths between the two domains and identifying the optimal target site lengths. Once these architectures were defined, TAL effector DBDs that recognise the HIV LTR sequences flanking the 16-bp HIV recombinase target site were then assembled with the selected HIV-excising catalytic domain variants to generate HIV TALERS. These HIV TALERS specifically target and promote the excision of a mock HIV-1 substrate in an in vitro recombination assay. This research provides a solid proof-of-concept for the use of CRs to target divergent novel target sequences, expanding their applicability for applied genome editing and wider biotechnological applications.

Table of Contents

Abstract	iii
Table of Contents	iv
Acknowledgements	vii
Abbreviations	ix
Chapter 1: Introduction	1
1.1 Synthetic Biology	2
1.1.1 Biomedical Synthetic Biology	2
1.2 HIV	4
1.2.1 HIV: Genome Structure	4
1.2.2 HIV-1: lifecycle.....	8
1.2.3 HIV-1: Therapy	9
1.3 Genome Editing	11
1.3.1 Genome Editing: Nuclease-based tools	12
1.3.1.1 Zinc finger nucleases (ZFNs).....	17
1.3.1.2 Transcription activator-like effector nucleases (TALENs).....	20
1.3.1.3 RNA-guided engineered nucleases (RGENs) - CRISPR.....	26
1.3.2 Genome Editing: HIV-1	27
1.3.2.1 Disruption of Host Cellular Receptors.....	27
1.3.2.2 Proviral mutation or excision.....	28
1.3.3 Genome Editing: Limitations of nuclease-based systems	29
1.4 Site-Specific Recombination	30
1.5 Tyrosine Recombinases	32
1.5.1 Tre and Brec1: reprogrammed tyrosine recombinases for HIV-1 excision	34
1.6 Serine Recombinases	38
1.7 The structure of Tn3 resolvase	42
1.8 Hyperactive (Deregulated) mutants of serine recombinases	46
1.9 Chimaeric Recombinases	48
1.9.1 Zinc Finger Recombinases	49
1.9.2 TALE Recombinases.....	51
1.9.3 RNA-guided Recombinases.....	54
1.10 Research Aim	55
Chapter 2: Materials and Methods	58
2.1 Bacterial Strains	59
2.2 Chemicals	59
2.3 Bacterial growth media	59
2.4 Antibiotics	60
2.5 Custom DNA synthesis	60
2.6 Plasmids	60
2.7 Plasmid Cloning	80
2.7.1 Annealing oligonucleotides	80

2.7.2	Restriction endonuclease digestion of DNA	80
2.7.3	Gel electrophoresis (I) - Agarose gel electrophoresis	81
2.7.4	Extraction of DNA from gel fragments	82
2.7.5	Ligation of DNA restriction fragments.....	82
2.7.6	Ethanol precipitation of DNA.....	84
2.7.7	Preparation of competent <i>E. coli</i> cells.....	84
2.7.8	Transformation of <i>E. coli</i> cells.....	85
2.7.9	Preparation of plasmid DNA	86
2.8	Estimating DNA concentration by UV spectrophotometry.....	86
2.9	DNA Sequencing	86
2.10	Plasmid design and construction	86
2.10.1	Expression plasmids.....	87
2.10.2	Substrate plasmids	88
2.11	<i>In vivo</i> recombination reactions - “MacConkey agar assay”	95
2.12	Protein Expression and Purification	98
2.12.1	Large scale induction of CR variants.....	98
2.12.2	Extraction and purification of His-tagged CRs	98
2.13	Estimating protein concentration	101
2.14	Gel electrophoresis (II) - Polyacrylamide gel electrophoresis	103
2.14.1	Discontinuous polyacrylamide gel electrophoresis.....	103
2.14.2	Denaturing PAGE for oligonucleotide purification	107
2.14.3	Native PAGE for fluorescent electrophoretic mobility shift assay (fEMSA) 108	
2.15	<i>In vitro</i> binding reactions.....	108
2.16	<i>In vitro</i> recombination reactions.....	109
2.17	Computer/Software	110
Chapter 3: Engineering the N-terminal catalytic domain.....		113
3.1	Research Strategy: Modular Engineering	114
3.2	Introduction: Engineering the Catalytic domain	116
3.3	Results.....	118
3.3.1	Target site selection	118
3.3.2	TATA-CR selection target substrate design	122
3.3.3	Engineering active TATA-ZFRs	124
3.3.4	Strategy 1: Analysis of broadened-specificity mutants.....	124
3.3.5	Strategy 2: Screening of Tn3toSin E-helix libraries	129
3.3.5.1	Recombination activity of 1st generation mutants on target sites	133
3.3.6	Strategy 3: Screening of hyperactive mutants on HIV sites.....	137
3.3.7	Strategy 4: Designing active TATA-CR catalytic domains based on position -3 .	142
3.3.7.1	Analysis of CR_TATA left core sequence.....	142
3.3.7.2	Analysing position -3 nucleotide preference of Tn3 Z-site	143
3.3.7.3	Library selection for Tn3 ZFR mutants with activity on altered position -3 substrate plasmids	146
3.3.8	Rational design of new mutants	151
3.3.9	Optimizing activity of working mutants	154
3.4	Discussion.....	156
3.4.1	Determinants of recombination on non-cognate sites.....	156
3.4.2	Mutational analysis	157

<u>Chapter Four: Defining the TALER architecture</u>	<u>165</u>
4.1 Introduction	166
4.2 Results.....	174
4.2.1 <i>In vivo</i> properties of TALER	174
4.2.2 <i>In vitro</i> structural optimization of TALER activity	176
4.2.3 <i>In vitro</i> characterization of TALER activity	187
4.2.4 Binding Properties of TALER variants	201
4.3 Discussion.....	206
4.3.1 <i>In vivo</i> activity of TALERS.....	206
4.3.2 Structural definition of TALER architecture	207
4.3.3 T-site architecture.....	208
4.3.4 TALER catalytic domain as main driver of recombination	209
4.3.5 TALER activity and recombination product distribution	209
<u>Chapter Five: Construction and Characterization of TATA-TALER</u>	<u>212</u>
5.1 Introduction	213
5.2 Results.....	213
5.2.1 <i>In vitro</i> analysis of the activity of the selected mutant catalytic domain.....	213
5.2.2 Design of the HIV CR_TATA TALE DBDs	218
5.2.3 Construction of the full HIV CR_TATA TALERS	228
5.2.4 Characterization of the activity of HIV CR_TATA TALERS.....	228
5.3 Discussion.....	238
5.3.1 TALERS demonstrate significant programmable targeting capacity	238
5.3.2 HIV TALERS specifically target CR_TATA_target site	238
<u>Chapter Six: Conclusions.....</u>	<u>240</u>
6.1 Biotechnological implications of programmable CRs	241
6.2 Improving TALER activity: Structural Considerations	243
6.3 Improving TALER activity: Technical Considerations.....	243
6.4 Outlook: Genome editing for proviral DNA excision.....	244
<u>Bibliography</u>	<u>246</u>
<u>Appendix.....</u>	<u>259</u>

Acknowledgements

I would like to thank the College of Science and Engineering, University of Glasgow and the Engineering and Physical Sciences Research Council (EPSRC) for providing the generous funding that supported this work. I appreciate the University of Ilorin for granting me a study leave. I am also grateful to Jon Cooper and Julien Reboud, my supervisors in the School of Engineering for the opportunity to do this research and for trusting me to embark on work outside the confines of the Rankine building. Thank you for approving funding for all the conferences I attended and spoke at as well.

I would like to applaud my Bio-Supervisor, Marshall, who found almost every missing semi-colon in the previous drafts of this thesis. Thanks for all the support and advice throughout these few years and for allowing me let loose in the lab while I explored all forms of tangential ideas. And here is a dose of thanks to Femi Olorunniji for daily support and guidance and for trying to be enthused every time I came with my famous 'I've been thinking...' and 'I have a quick question' lines. He attempted to convert me into a biochemist, I am not quite sure how that venture turned out. I must say though that I am a better, more rational and focused scientist because of your mentorship. I also appreciate our super-technician, Arlene, for all her help and patience and to Chris Proudfoot, who I never met, but who laid the groundwork for some of the work here. A lot of thanks to everyone in the Hooker lab that I have worked with over the years - to Steph, Rich, Hayley, Al, Gill, James, Anna, Emanuele, Steve, Jia, Gillian and to Sean, Sally, Martin and others for their helpful support, encouragement and for a great working environment during this project.

And to my family, thank you for being a constant haven for me, albeit from a huge distance. I am grateful for all my friends, you are a blessing. I am also thankful for Dr Anibijuwon, Dr Kolawole and all my colleagues at the University of Ilorin. Thank you Jide Oyewole, for your continued support and comfort.

I declare that, except where explicit reference is made to the contribution of others, this thesis and the work presented in it are my own, generated as the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Jumai Adeola Abioye

Abbreviations

Units

k:	10^3	°C:	degrees Celsius
c:	10^{-2}	mol:	moles
m:	10^{-3}	M:	Molar
μ :	10^{-6}	rpm:	revolutions per minute
n:	10^{-9}	V:	Volts
p:	10^{-12}	hr:	hours
g:	grams	mins:	minutes
m:	metres	kb/kbp:	kilo base pairs
L:	litres	OD:	optical density

Commonly-used abbreviations

A:	Adenine	PAGE:	Polyacrylamide gel electrophoresis
aa:	Amino acid	PDB:	Protein Data Bank
APS:	Ammonium persulphate	RGEN:	RNA-guided engineered nuclease
ART:	Antiretroviral therapy	RNA:	Ribonucleic acid
bp:	Base pair	RVD:	Repeat variable di-residues
kb:	Kilobases	SDS:	Sodium dodecyl sulphate
C:	Cytosine	SSR:	Site-specific recombinase
CAS:	CRISPR- associated genes	T:	Thymine
CR:	Chimaeric recombinase	TAE:	Tris-acetate-EDTA
CRD:	Central repeat domain	TALE:	Transcription activator-like effector
CRISPR:	Clustered regularly interspaced short palindromic repeat	TALEN:	TALE nuclease
DBD:	DNA-binding domain	TALER:	TALE recombinase
DNA:	Deoxyribonucleic acid	TEMED:	Tetramethylethylenediamine
DTT:	Dithiothreitol	Tn3NM:	Activated Tn3 resolvase mutant NM contains mutations R2A, E56K, G101S, 102Y, M103I and Q105L
EDTA:	Ethylenediaminetetraacetic acid	Tris:	Tris(hydroxymethyl)aminomethane
G:	Guanine	UV:	Ultraviolet
GFP:	Green fluorescent protein	WHO:	World Health Organization
HIV:	Human immunodeficiency virus	ZFN:	Zinc finger nuclease
LANL:	Los Alamos National Laboratory	ZFR:	Zinc finger recombinase
LTR:	Long terminal repeat		
NTR:	N-terminal region		

Chapter 1: Introduction

1.1 Synthetic Biology

Synthetic biology repurposes life and/or its principles. This emerging discipline brings together knowledge and skills from genetics, molecular biology, engineering, physics, computational sciences and social sciences among others to create and (re-)design biological parts, tools and systems. In summary, synthetic biology is the engineering of biology for useful purposes (Khalil and Collins, 2010; Serrano, 2007). This could mean repurposing naturally existing biological components for new applications through reassembly and optimization or the creation of new-to-nature systems and artificial metabolic pathways based on biological principles and design (Agapakis, 2014). The outputs of synthetic biology find use in environmental applications, agriculture, medicine and many other fields.

1.1.1 Biomedical Synthetic Biology

Global health currently poses unique challenges. The rise of antimicrobial resistance could be leading us into a post-antibiotic era, climate change is driving neglected tropical diseases into previously uninfected terrains, and an increasingly ageing population is set to have profound economic consequences across the world (WHO, 2017). Synthetic biology holds several solutions for these challenges; the field aims to provide technical tools needed to drive low-cost innovation for diagnosis, prevention, treatment and even cure in human health (van Passel *et al.*, 2014).

Advances in genomics, DNA production and high-throughput screenings allow synthetic biologists to tackle major medical issues with increasing precision and speed. A key tool for biomedicine is the (re-)design of novel biological sensors that can detect and report the presence of important target analytes. By optimising earlier systems such as the glucose-oxidase-based biosensors used in glucometers, newer biological sensors comprise complex integrated pathways in whole cells or cell-free toolkits (Alhadrami, 2017). The ability of living systems to amplify signals ensures that previously undetectable levels of toxic metabolites can now be quantified. These biosensors find use in the rapid diagnosis of infectious diseases

and environmental monitoring for disease prevention (Vidic *et al.*, 2017; Sin *et al.*, 2014; French *et al.*, 2011).

The integration of diagnostic biosensor technologies with microfluidic platforms allows the delivery of point-of-care diagnostics, an approach poised to reduce antimicrobial misuse and the spread of infections. In a different approach, the coupling of biosensors with genetic switches that can trigger responses to specific molecules holds unique promise for disease treatment. Zheng and colleagues (2017) recently reported the design and use of an engineered *Salmonella typhimurium* that overexpresses flagellin to elicit tumour-killing activity in mice. Disruptive approaches such as this demonstrate the potential of synthetic biology to improve the current standards and approaches in disease treatment (Felgner *et al.*, 2016).

Many of these ideas are moving from basic biological bench research into biopharmaceutical and industrial applications and are beginning to create a robust bioeconomy to maximize these discoveries for human use. Strain and metabolic engineering for cheaper and improved drug production, *de novo* protein design for drug delivery, and the design of microbial communities for gut disease prevention or cure are some other biomedical synthetic biology interventions to look out for.

This research work focuses on the use of synthetic biology for tackling a globally relevant problem and the optimization of a next-generation approach to personalized medicine - genome editing.

1.2 HIV

Of huge global biomedical significance are the pandemic infection, disease and complications caused by the Human Immunodeficiency Virus (HIV). There are currently over 36 million people living with the HIV infection globally (WHO, 2017). The virus attacks and invades the immune system of its host and over time the infection progresses into a full-blown disease, Acquired Immunodeficiency Syndrome (AIDS), where the patient's immune system is severely weakened, and the patient is susceptible to several life-threatening illnesses.

Genetic variability has led to the classification of two identified types of HIV: HIV-1 and HIV-2. Of these two, HIV-1 is more virulent and is the type responsible for the HIV pandemic. HIV-1 is also grouped into four groups, one major group (M) and three minor groups (N, O and P) (Hiv.lanl.gov, 2017). HIV-1 Group M is subdivided into 10 groups- A, B, C, D, F, G, H, J, K and the circulating recombinant forms (CRFs). CRFs are unique recombinant viruses formed from recombination events between two or more viruses of subtype A to K and which have infected three or more unrelated persons. There are currently 90 CRFs in the HIV Los Alamos National Laboratory (LANL) HIV database, the key source of global HIV genome sequence data.

1.2.1 HIV: Genome Structure

HIV is a member of the genus *Lentivirus* in the family *Retroviridae*. Lentiviruses are single-stranded enveloped RNA viruses that obligatorily integrate their reverse-transcribed DNA into the chromosome of their host cell as part of their lifecycle. They also have long incubation periods and can infect inactive cells.

HIV as an infectious particle (virion) contains two identical single strands of HIV RNA embedded in the viral core (Barré-Sinoussi *et al.*, 2013) (Fig. 1.1). The viral core also contains the proteins required for viral genome replication and is surrounded by matrix proteins and a viral envelope. The virion cannot replicate on its own; it is transmitted through bodily fluids including blood, semen, breast milk, vaginal fluid and anal fluid into host cells. The reverse-transcribed DNA

genome of HIV is about 9800 bases long and encodes three classes of genes: major structural genes (*gag*, *pol* and *env*), regulatory genes (*Tat* and *Rev*) and accessory genes (*Vif*, *Nef*, *Vpr* and *Vpu* or *Vpx*) (Fig. 1.2).

The *gag* gene, or group structural antigen, encodes the Gag precursor protein, p55 which is processed during maturation to make four smaller proteins - the matrix proteins (p17), capsid proteins (p24), nucleocapsid proteins (p7) and the late-assembly domain (p6) (GAC blood, 2016). These proteins assemble with the RNA to form the inner structure of the virion. The *pol* gene, usually expressed as a Gag-Pol precursor fusion protein (p160), is processed to four enzymes- the protease (p10), reverse transcriptase (p50), RNase H (p15), and integrase (p31). These enzymes have significant activities in the lifecycle of the virus (Fig. 1.3). The Env polyprotein (gp160), encoded by *env*, is cleaved to generate the envelope glycoproteins- the transmembrane proteins (gp41) and the surface proteins (gp120). Regulatory proteins, *Tat* and *Rev*, are essential for transcriptional activation and RNA splicing regulation respectively. The accessory proteins function in viral replication and in ensuring viral infectivity and virulence. HIV-1 encodes *Vpu* while HIV-2 encodes *Vpx*.

Flanking both ends of the viral coding region are identical sequences of 640bp each. These regions called the long terminal repeat (LTR), are generated during the reverse transcription of the HIV-1 viral ssRNA (Fig. 1.1). They are involved in the integration of HIV into the host genome and contain several binding sites for host and viral factors. The HIV-1 LTR has been significantly implicated in disease virulence and pathogenesis (Hiv.lanl.gov, 2017; Krebs *et al.*, 2002). LTR activity is predicted to deregulate host systems and increase virulence by modulating viral lifecycle and cytotoxicity. Although having an identical sequence, the two LTRs are functionally different. The 5' LTR serves as the core promoter for the integrated virus while the 3' LTR functions in RNA polyadenylation.

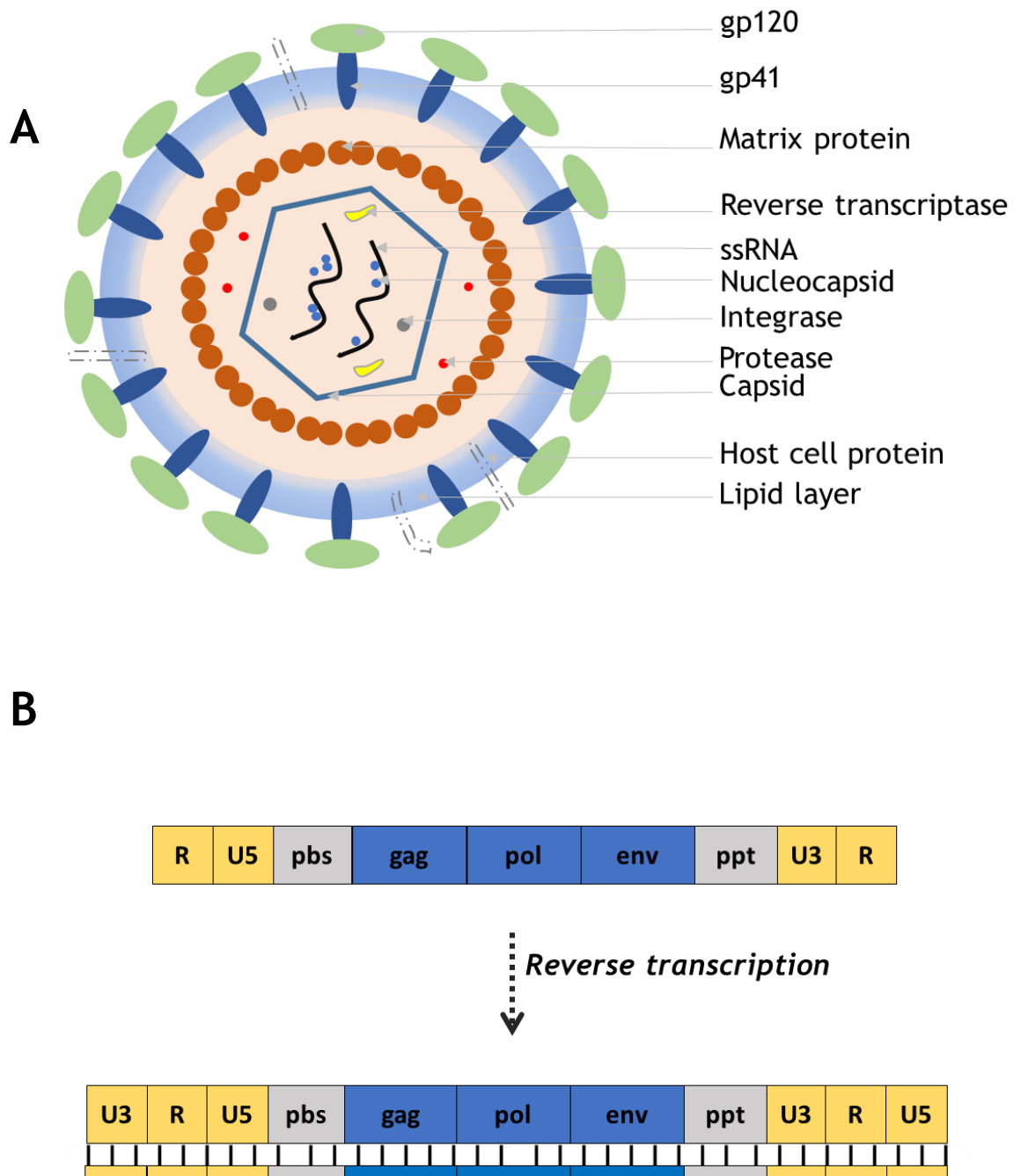


Figure 1.1: The structure of HIV. **A.** Simplified view of HIV Virion. The HIV viral particle contains two copies of positive-sense ssRNA surrounded by a capsid and matrix proteins and encapsulated in a lipid envelope. The virion also contains important viral enzymes such as the integrase and reverse transcriptase. **B.** Reverse transcription of ssRNA to DNA. The LTRs flanking the coding region are generated from a single copy on the ssRNA during a multi-step reverse transcription process. They serve as key control centres for gene expression in the provirus. Each LTR is split into three regions- U3, R and U5. The LTR regions are shown in yellow, HIV genes in blue, and two retroelements involved in the process, pbs (primer binding site) and ppt (polypurine tract), in grey.

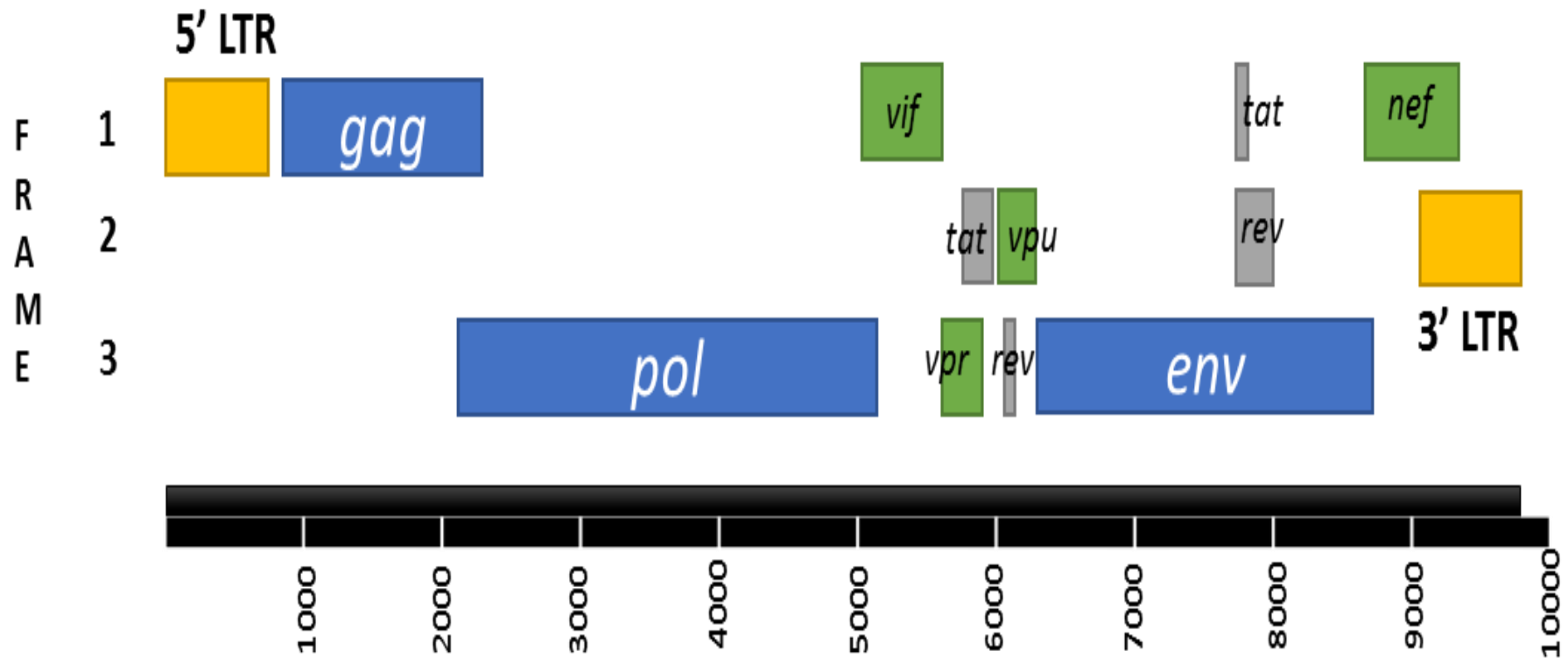


Figure 1.2: HIV-1 gene map. HIV-1 DNA spans almost 10 kbp in length and encodes several genes that enhance viral virulence. Alternative RNA splicing leads to the generation of different gene products, maximizing the viral genome space. The provirus is flanked at both ends by a 640-bp direct repeat sequence called the long terminal repeat (LTR). The other genes in HIV-1 encode structural, regulatory and accessory proteins (based on HIV-1 HXB2 genome with GenBank accession number K03455). The ruler provided is graduated in base pairs (bp) (Adapted from Hiv.lanl.gov, 2017).

1.2.2 HIV-1: lifecycle

Upon infection, the virus enters target immune cells such as T-cells and macrophages by binding to host cellular surface receptors and fusing to the cell using its fusion protein, gp41 (Fig. 1.3). The viral capsid is then uncoated, releasing viral RNA and enzymes into the host cell. The viral reverse transcriptase hijacks host cell machinery to create viral complementary DNA (cDNA) from the ssRNA. This cDNA, assembled with other viral and cellular components into a 'viral pre-integration complex' (PIC), is then transported to the nucleus and integrated into the cellular host genome using viral integrase.

The integrated viral DNA is called the provirus. Using the host machinery, transcription and translation of proviral DNA results in long chains of viral protein (polyprotein) and ssRNA, the building blocks of new HIV viruses. These polyproteins and RNA assemble at the cell surface and bud off to form immature viruses. Within each immature virus, viral protease cleaves the long chains of protein to form smaller proteins that combine to form the mature, infectious HIV virion.

Because of the intrinsic nature of the viral pathology (genomic integration) and the importance of the target cells to the host (immune cells), there is currently no cure for the viral infection. As the viral DNA is integrated into the host genome, killing the provirus might also kill the infected cell. In addition to this, HIV-1 has a high genetic variability. A short lifecycle and an error-prone reverse transcriptase result in a high degree of mutability and a strong ability to elude many therapeutic approaches, yielding several multi-drug resistant HIV-1 subtypes.

1.2.3 HIV-1: Therapy

The current approved therapeutic methods attempt to avoid multi-drug resistance through a combination therapy, previously referred to as highly active antiretroviral therapy (HAART) or combination antiretroviral therapy (cART) but now simply known as antiretroviral therapy (ART). ART presents multiple obstacles to HIV-1 replication, typically by targeting viral proteins such as integrase, reverse transcriptase, protease and fusion proteins (Fig. 1.4).

For example, nucleoside/nucleotide reverse-transcriptase inhibitors (NRTIs) and non-nucleoside reverse-transcriptase inhibitors (NNRTIs) disrupt the ability of viral reverse transcriptase to generate full-length viral cDNA (Arts and Hazuda, 2012). NRTIs are analogues of natural DNA molecules and competitively inhibit reverse transcriptase activity by prematurely terminating DNA incorporation, because of a lack of the 3'-OH required for the formation of a phosphodiester bond with the next molecule. Unlike NRTIs, NNRTIs inhibit reverse-transcriptase activity by binding directly to reverse transcriptase to induce conformational changes that render the protein inactive or defective.

While ART strategies work against active HIV-1-infected cells and have been used effectively to suppress plasma viral load to below detectable levels in infected persons, the integrated proviral DNA is unaffected (Sarkar *et al.*, 2007; Wong *et al.*, 1997). This stably integrated proviral DNA is passed on to host cell progeny; therefore, life-long treatment is usually required to manage the infection. Transcriptionally inactive states of the provirus can be harboured in resting memory CD4⁺ T cells and have the capacity to stay silent for years in this viral reservoir (Siliciano and Greene, 2011). These infected cells serve as latent reservoirs of the infection, preventing the eradication of the virus. Active infection can also be triggered by a change in the infected person's immunological state, such as cessation of treatment (Chun *et al.*, 2007). This implies that current ART strategies are incapable of 'curing' the HIV infection.

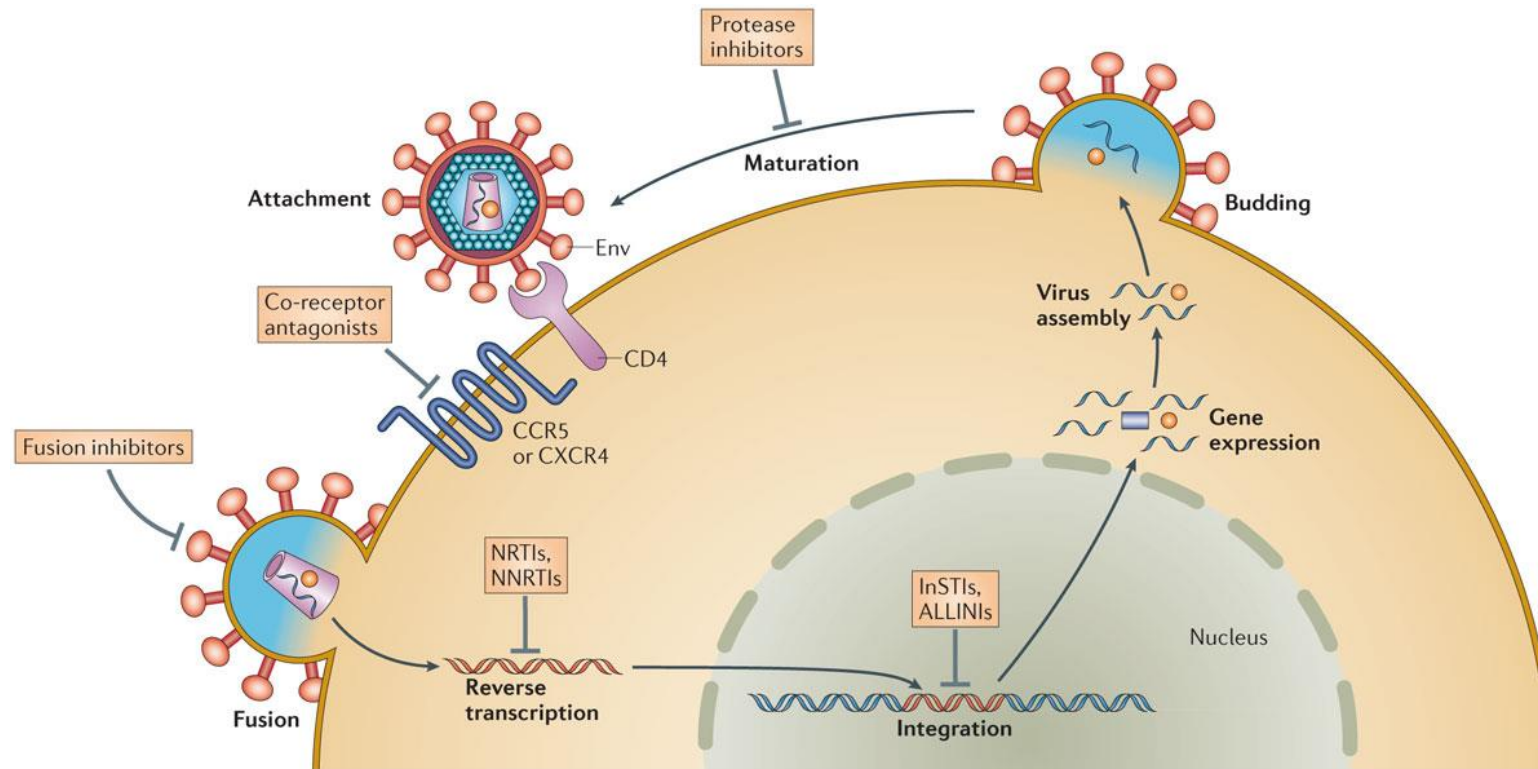


Figure 1.4: ART drug targets. Some of the proteins involved in the HIV-1 life-cycle include fusion proteins, reverse transcriptase, integrase and proteases. These proteins and the key processes of the HIV life cycle are the current targets of several antiretroviral therapy strategies. Fusion and entry inhibitors impair the ability of the virus to infect new host cells; NRTIs and NNRTIs target the viral reverse transcriptase using different mechanisms to prevent cDNA generation; integrase inhibitors prevent the integration of translocated cDNA into the host genome and protease inhibitors prevent the production of mature infectious virions by preventing the proteolytic activity of viral protease on the polyprotein (originally from Laskey and Siliciano, 2013).

In addition to HIV-1 infection latency, other issues with ART include; drug toxicities and off-target effects, lifelong burden of compliance to treatment regimens by patients, and increasing cases of HIV-1 drug resistance (Ahlenstiel *et al.*, 2015; Hsu *et al.*, 2013). The cost of lifelong ART is also not sustainable for the most affected. Most people living with HIV reside in low- and middle-income countries with almost 70% of them in Africa in 2016 (WHO, 2017).

An alternative strategy is thus necessary to tackle HIV. A cure should effectively eradicate the provirus from all cells of the infected person. Approaches proffered for clearing the provirus include the activation of all latent reservoirs for ART or the excision of the provirus from memory T- cells. Genome editing strategies could be designed to achieve these.

1.3 Genome Editing

Genome editing has recently gained prominence for research into the genetic basis of diseases, as well as for novel therapeutic purposes and strain improvement for industrial bioprocesses (Cox *et al.*, 2015). Its approaches can be used to tackle diseases affecting the human genome like the HIV-1 infection.

Genome editing involves the use of natural and engineered enzymes as ‘molecular scissors’ for targeted DNA manipulation such as excision, integration, exchange and rearrangement. These systems can be used to permanently replace defective genes, introduce new functions or excise unwanted genetic regions. The most prominent tools in the current approach to genome editing are nuclease-based platforms. For example, these enzymes have been used for site-specific genomic alterations in the generation of disease models (Ding *et al.*, 2013), epigenetic gene regulation (Hilton *et al.*, 2015), generating HIV-1 resistant CD4(+) T cells (Perez *et al.*, 2008), driving human genomic transgenesis (Dekelver *et al.*, 2010) and yellow fever mosquito (*Aedes aegypti*) genome modification (Dong *et al.*, 2015).

1.3.1 Genome Editing: Nuclease-based tools

Nucleases are phosphodiesterases that cleave the phosphodiester bonds in DNA or RNA molecules (Yang, 2010). Exonucleases cleave at the end of a polynucleotide chain, thereby removing one nucleotide at a time while endonucleases cleave in the middle of a polynucleotide chain. In nature, nucleases function as regulators of high-fidelity cellular processes such as replication, RNA processing and programmed cell death. They are either sequence-specific or structure-specific. Some sequence-specific endonucleases called restriction enzymes find use as routine molecular biology tools for gene cloning and analysis.

Nucleases, such as the FokI endonuclease, have been engineered for use in targeted genome engineering to cut defined genomic nucleic acid sequences. Engineered nucleases serve as 'artificial' restriction enzymes responsible for recognizing, binding and then induction of single-strand or double-strand breaks on the target site.

There are three major systems of engineered nucleases (Kim and Kim, 2014). Two of them are generated by fusing nuclease domains with altered or no sequence specificity to programmable DNA-binding protein domains - zinc finger (ZF) proteins and transcription activator-like effector (TALE) proteins - to generate Zinc Finger Nucleases (ZFNs) and Transcription Activator-Like Effector Nucleases (TALENs). The third type of nucleases are RNA-guided engineered nucleases (RGENs) such as the clustered regularly interspaced short palindromic repeat (CRISPR)- Cas (CRISPR-associated) systems which use RNA (rather than protein domains) to target and guide a nuclease (e.g. Cas9) to specific DNA sequences (Fig. 1.5).

In the design of ZFNs and TALENs, the FokI endonuclease domain is usually fused to the C-terminal end of the programmable DNA-binding proteins. FokI endonuclease is a Bacterial type IIS endonuclease with functionally separable domains (Durai *et al.*, 2005). Li, Wu and Chandrasegaran (1992) reported that the

architecture of the endonuclease allowed the physical separation of protein into distinct and functional DNA binding domains (DBDs) and cleavage domains. The cleavage domain has no apparent DNA sequence specificity, and this allows the retargeting of this domain to any site for cleavage by programming the DBD appropriately. Dimerization is required for the activity of the endonuclease domain at the target site, and this property is usually counted on to enhance nuclease specificity. Several approaches have been taken to optimise the activity and specificity of the FokI engineered nucleases (Sakuma and Woltjen, 2014). A monomeric nuclease domain variant, the T_{ev}I homing endonuclease, has also been utilised in generating engineered nucleases (Kleinstiver *et al.*, 2014).

RGENs are the most recently discovered of the nuclease-based genome-editing tools. For the most commonly used RGEN, the CRISPR-Cas9 system (here, the *Streptococcus pyogenes* CRISPR-Cas9 system is considered), the Cas9 nuclease domain is guided to the cleavage target site by short guide RNAs where it induces a double strand break (Doudna and Charpentier, 2014) (Fig. 1.5). Cas9 sequence specificity is conferred by the guide RNA and a short sequence on the host genome ('NGG', where N is any 'nucleobase'), the protospacer-associated motif (PAM). Two variants of wild-type Cas9 have been generated; a Cas9 nickase (Cas9D10a) that only generates single-strand breaks, and a nuclease-deficient Cas9 (dCas9) that can be used for transcriptional regulation.

Engineered nucleases do not catalyse re-joining or ligation of the cut ends of the DNA. Once the target site is cut, lesion repair is usually modulated by host cell DNA damage repair mechanisms (Cox *et al.*, 2015). Repair is either in the form of non-homologous end joining (NHEJ), an error-prone and non-specific process, or homology-directed repair (HDR) where an accompanying repair template is introduced for precise genome editing (Fig. 1.6). NHEJ usually re-joins the two ends, introducing varying lengths of insertions and/or deletions (indels). These indels can lead to frameshift mutations in target genes which can destabilize cellular processes. However, where gene inactivation or suppression is the expected therapeutic outcome, this could be desirable. With HDR, since the repair

outcome is predefined by the template, therapies involving gene correction, replacement, activation and insertion can be designed.

The next few sections will elaborate on the three nuclease-based genome editing tools with a focus on the DNA-binding modules. Since effector domains such as epigenetic modifiers, transposases, transcriptional effectors and recombinases can be fused to these DBDs for targeting the genome for modification, it is important to discuss them sufficiently. RGENs (the CRISPR- Cas9 system) will only be briefly discussed in this work. Reviews by Ding *et al.*, (2016) and Wang *et al.*, (2016) provide ample information on RGENs.

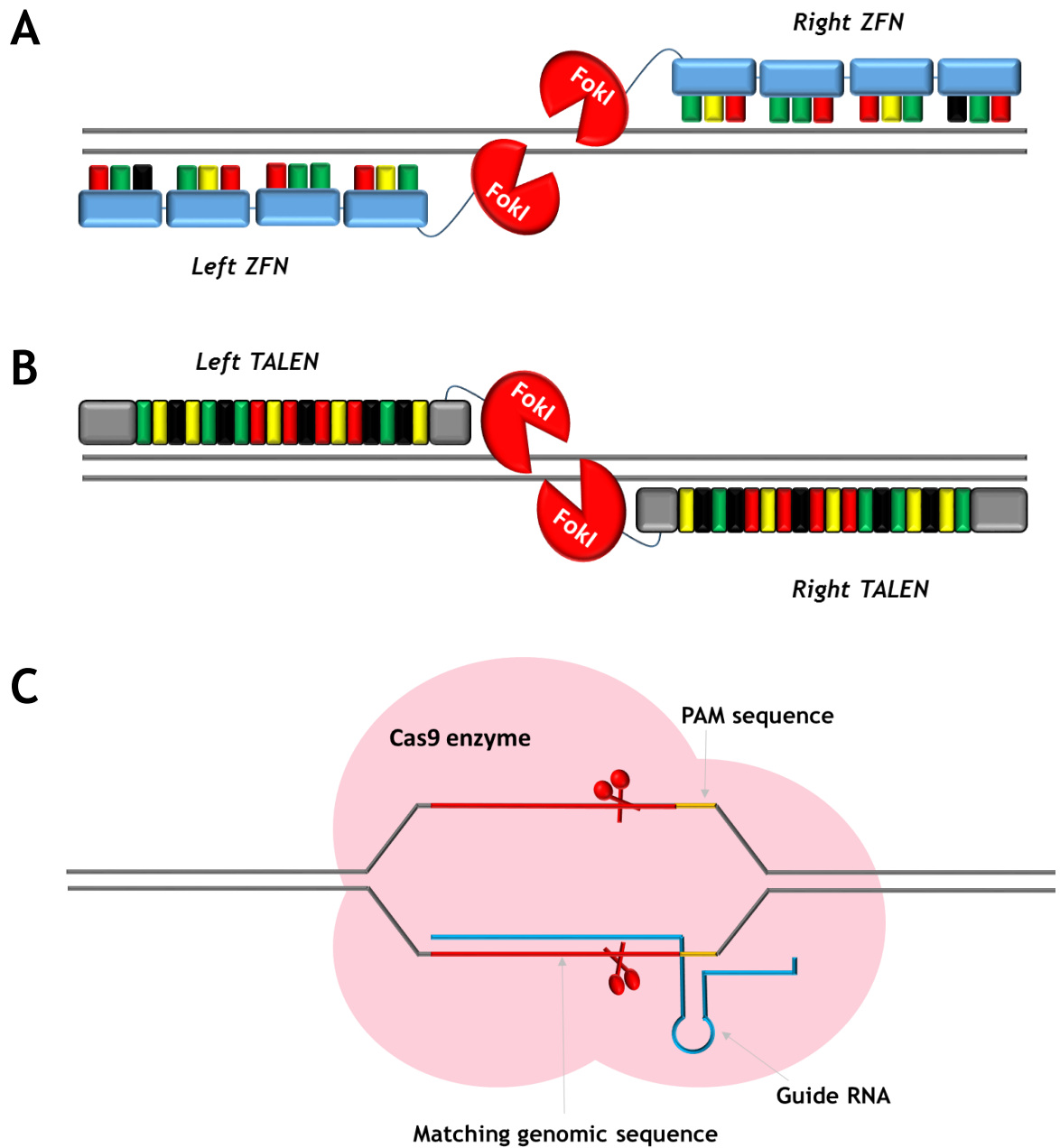


Figure 1.5: Nuclease-based genome editing tools. A. Zinc finger nucleases. **B.** Tal-effector nucleases. **C.** CRISPR/Cas9 system. ZFNs and TALENS usually function as dimers for inducing double-strand breaks at the target site using a FokI nuclease domain. Zinc fingers recognize DNA triplets while TALEs demonstrate a one-to-one association with nucleotides at their target sites. In contrast, the native CRISPR/Cas9 system is RNA-guided and functions as a monomer, inducing double-strand breaks at the target site with two nuclease domains in the Cas9 protein. The target site sequence is specified by a guide RNA and is flanked at the 3' end by a 3-bp motif sequence, PAM.

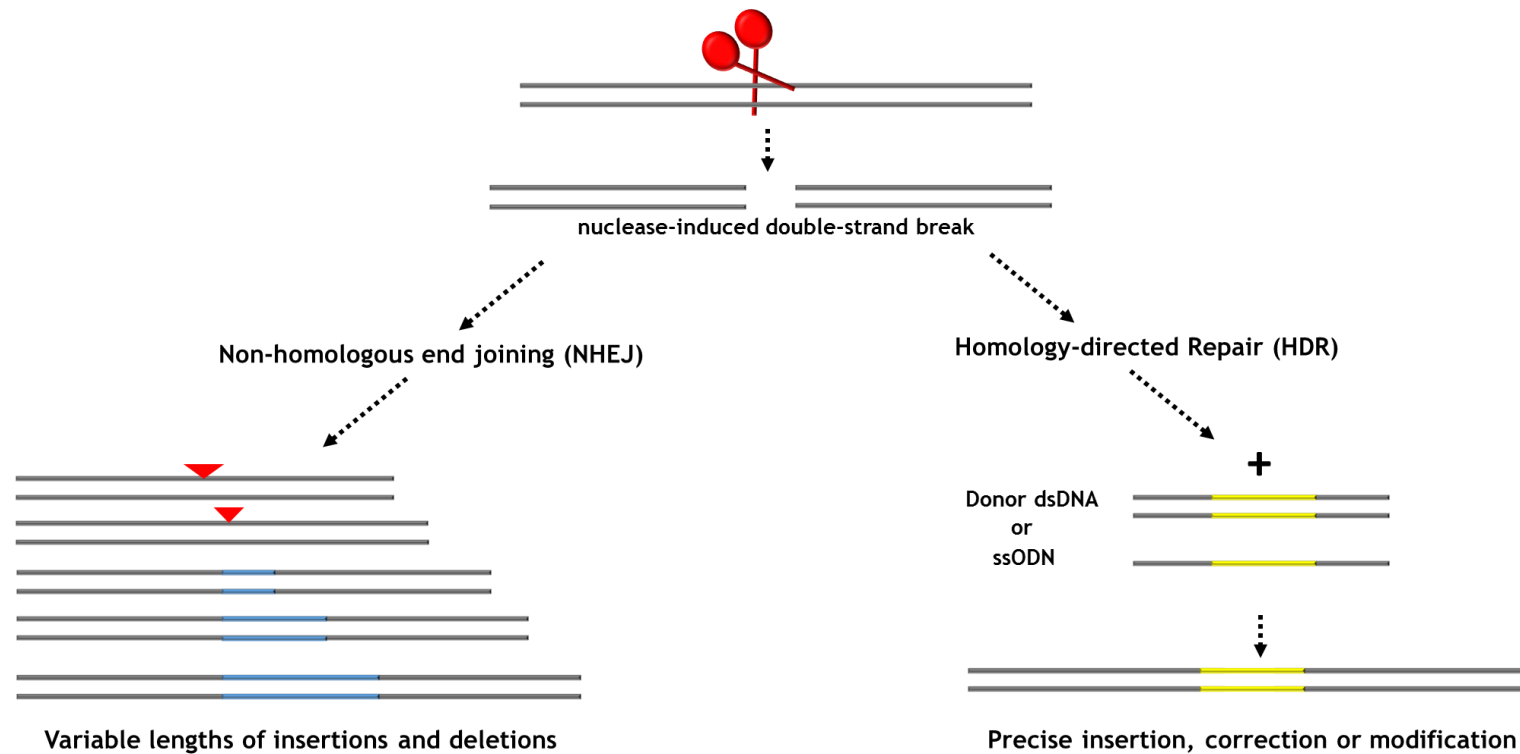


Figure 1.6: Nuclease-based genome editing strategies and repair mechanisms. Double strand breaks induce two different repair pathways in cells. Non-homologous end joining (NHEJ) religates the cut DNA ends without the requirement for a template strand and results in insertions and deletions (indels) which can lead to fatal genetic mutations. The outcomes of the NHEJ pathway are very variable with indels ranging from 1- to 200-bp. In cases where gene knockouts are the intended therapeutic outcomes, this system is quite useful. However, for more control, the homology-directed repair (HDR) pathway can use an accompanying homologous template DNA or a single-strand oligodeoxynucleotide (ssODN) to generate precise repair outcomes.

1.3.1.1 Zinc finger nucleases (ZFNs)

Zinc finger (ZF) proteins are a large family of small, modular proteins that contain one or more stabilizing zinc ions. Most bind nucleic acids in a sequence-specific manner but a few of them target proteins and lipid substrates. They vary widely in structure and function. Their functions range from apoptosis regulation and chromatin remodelling to gene transcription and translation (Laity *et al.*, 2001).

The Cys2-His2 (C2H2) type was the first type of zinc finger protein to be discovered; it was found in the transcription factor TFIIIA of *Xenopus laevis*, the African clawed frog (Miller *et al.*, 1985). C2H2 zinc finger proteins adopt an invariable $\beta\beta\alpha$ structure holding a single zinc ion by a tetrahedral coordination (Fig. 1.7). Each C2H2 finger typically makes contact with a triplet nucleotide sequence, reading the major groove of the DNA. Several different motifs can be stacked continually to bind multiple nucleotide triplets, targeting a long stretch of nucleotides. Each finger is usually separated by a highly conserved 5-amino-acid residue 'linker' to provide optimum spacing for recognition and protein-DNA interaction (Wolfe *et al.*, 2000). A theory has been proposed that this linker also plays a major role in 'locking' the protein to its target sequence (Laity *et al.*, 2000; Wolfe *et al.*, 2000). The model posits that the zinc finger protein diffuses along the DNA until it encounters its target sequence, then a conformational change is induced with the linker capping the C-terminus of the preceding helix. With interesting similarity to the C-terminus of the E-helix region of Tn3 ($\gamma\delta$) resolvase, this linker is dynamically disordered in the absence of DNA and only forms structurally upon binding to DNA (Section 1.7).

There have been many attempts to deduce a qualitative ZF protein-binding code for the synthesis of novel and unique fingers (Liu and Stormo, 2008; Kaplan *et al.*, 2005). Several assays and computer programmes have been developed over the years to enhance the large-scale selection and synthesis of novel zinc fingers to recognize any nucleotide triplet (Maeder *et al.*, 2009; Sander *et al.*, 2011). These include the oligomerized pool engineering, OPEN, zinc finger design protocol (Maeder *et al.*, 2009) which can be used to generate several zinc finger proteins

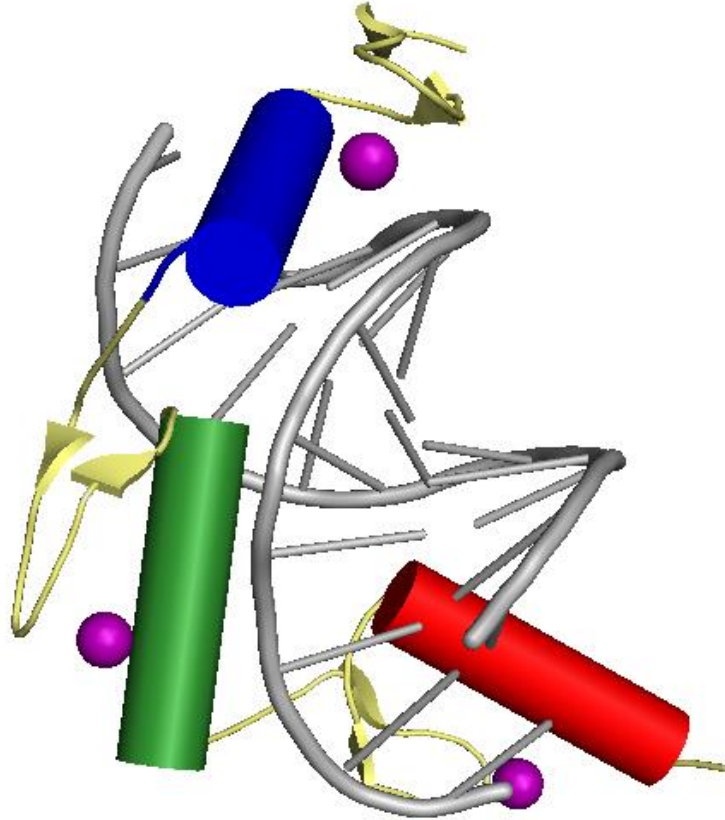
in 8 weeks and has been used successfully for the modular design and assembly of zinc finger domains that function in chimaeric systems such as zinc finger nucleases (Hermann *et al.*, 2012). Hundreds of thousands of fingers are combined in the pool. The protocol utilizes a selection design based on phage display and a bacterial-2-hybrid system in *Escherichia coli* for reporting zinc finger activity on a target recognition site.

An interesting feature of C2H2 zinc finger proteins that affects the selection and design of zinc finger proteins is the context dependency of DNA recognition by the zinc fingers. Basically, the specificity, efficiency and binding properties of each finger depend not only on its target sequence but also on the neighbouring sequences. This means that stacking multiple zinc fingers to target a stretch of sequence would not always yield the desired outcome (Liu and Stormo, 2008).

A protocol similar to OPEN, but selection-free and less resource-intensive, the context-dependent assembly method, (CoDA), attempts to tackle this issue by recombining already-characterized zinc finger proteins (CoDA units) with overlapping middle fingers to generate novel proteins (Sander *et al.*, 2011). The current archive of CoDA units (and in fact other ZFP pools) is expected to find a potential target site once in every 500 bp.

Zinc finger nucleases have been successfully used for genome editing in human cells, animals and plants, finding significant use in therapy (Chandrasegaran and Carroll, 2016). A ZFN application to HIV-1 genome editing is discussed in Section 1.3.2.

A



B

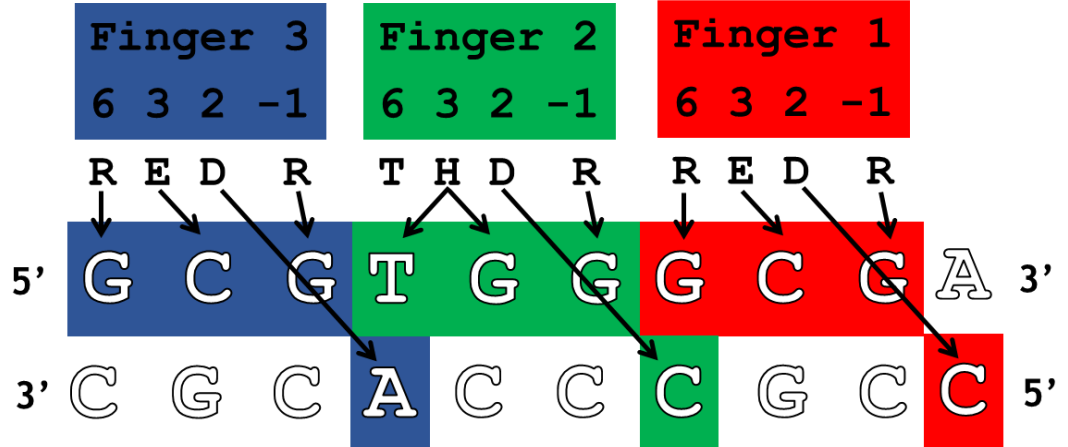


Figure 1.7. Zif268-DNA interactions. A. Cartoon representation of crystal structure of the 3- zinc finger protein, Zif268, bound to its DNA target site (PDB: 1AAY- Elrod-Erickson *et al.*, 1996). Zif268 is a modular protein that acts as a monomer. It is made up of three fingers. The fingers are shown as blue, green and red cylindrical helices. Zinc atoms are shown as purple spheres. B. Finger-DNA interactions. Each finger makes direct interactions with a nucleotide triplet and another nucleotide on the second strand adjacent to the target sequence of the neighbouring finger. As such, each finger is context-dependent on the other two fingers.

1.3.1.2 *Transcription activator-like effector nucleases* (TALENs)

TALEs are bacterial virulence factors secreted by the plant pathogens, *Xanthomonas sp.* to regulate plant gene expression for enhancing colonization and survival during infection (Bogdanove *et al.*, 2010). They enter the cytoplasm of plant cells through the bacterial type III secretion system, translocate to the nucleus and target plant gene promoters by mimicking eukaryotic transcription factors. TALE activities increase disease susceptibility of host cells and enhance the pathogenicity of *Xanthomonas sp.* causing growth defects and diseases such as hypertrophy and bacterial blight of rice.

TALEs are modular proteins which consist of a central DNA-binding tandem repeat domain (CRD) placed between an N-terminal region that contains the type III secretion signal (T3S) domain, and a C-terminal region with a nuclear localization signal (NLS) domain and a transcriptional activation domain (AD) (Gao *et al.*, 2012) (Fig. 1.8). The central tandem repeats consist of 33 to 35 amino acid residues, with variations usually only in two residues at positions 12 and 13 which are responsible for nucleotide recognition and are referred to as repeat-variable di-residues (RVD) (Sanjana *et al.*, 2012; Zhang *et al.*, 2011). The repeat closest to the C-terminal domain of the protein is usually truncated to 20 amino acids and is referred to as a 'half' repeat.

There are currently more than 20 known RVDs; however, 7 of them account for about 90% of known repeats (Mak *et al.*, 2012). Each RVD has specific association with one or more nucleotides; for example, the RVD NI (Asparagine-Isoleucine) binds preferentially to A, HD (Histidine-Aspartic acid) to C, NG (Asparagine-Glycine) to T and NN (Asparagine-Asparagine) to both G and A (Fig. 1.9). This one-to-one RVD-DNA interaction yields a code that enables the reconfiguration of TALEs for custom DNA recognition, providing a highly programmable tool for generating designer enzymes for targeting novel sites.

Naturally-occurring TALEs have repeats ranging from 1.5 to 33.5. However, research by Boch *et al* (2009) suggests that at least 6.5 repeats are required for TALEs to be functional. TALEs with 10.5 or more repeats showed strong reporter gene activation. Their recognition sequence is usually preceded by a 5'-thymine at position 0 (T^0) that seems to have a function in promoter activation. Mutation of this 5'-thymine has been shown to significantly reduce the activity of natural TALEs on their targets, although this effect is relatively minimal on artificial (designer) TALEs (Jankele and Svoboda, 2014; Meckler *et al.*, 2013).

Crystal structures of naturally-occurring and designer TALEs (dTALEs) in both DNA-free and DNA-bound forms have helped in the understanding of TALE structure and binding mechanisms and informed the design of functional fusion proteins of TALEs (Deng *et al.*, 2012; Gao *et al.*, 2012; Mak *et al.*, 2012). Mak *et al.*, (2012) presented the crystal structure of a 36-bp dsDNA-bound TALE, PthXo1 from rice pathogen, *Xanthomonas oryzae*. The protein construct has 23.5 repeats spanning residues 127 to 1149 of PthXo1. Each of the repeats forms a helix-loop-helix bundle; two α -helices from amino acid positions 3 to 11 and 14 to 33 with an interhelical loop (residues 12 and 13) containing the RVDs that coordinate DNA-specific interactions. The packing of tandem repeats forms a right-handed superhelical structure that wraps around the DNA major groove and the helices are arranged in a left-handed sequential packing.

The side-chain of the first RVD residue at position 12 makes intra-repeat backbone contacts with position 8 which stabilizes the RVD loop, while the second RVD (position 13) reaches into the major groove of the target DNA and makes direct nucleobase-specific interactions with the corresponding base on the DNA sense strand (Fig. 1.8). Seven different types of RVDs were observed in PthXo1- HD, NG, HG, NG, NI, NN and N* (missing a residue at position 13). The observed base-specific interactions are based on van der Waals contacts and hydrogen bonds between the RVD residue at position 13 and the bases. Other non-specific interactions are observed at positions 16 and 17 of each repeat.

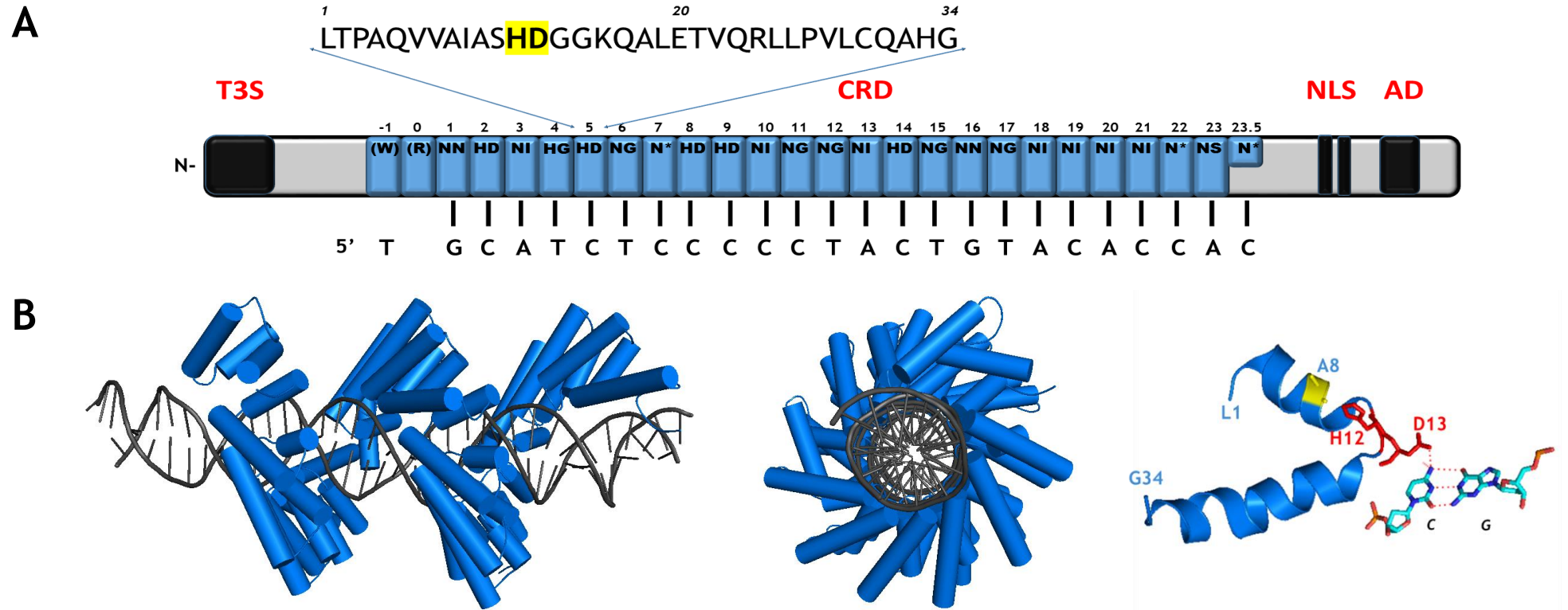


Figure 1.8: Domain organization and crystal structure of wild-type PthXo1. **A.** The domains of PthXo1 are organized from N-terminal to C-terminal into a type III secretion signal (T3S) domain, the central repeat domain (CRD), a nuclear localization signal (NLS) domain and a transcriptional activation domain (AD). The CRD has 23.5 repeats shown in the crystal structure image. The RVD in each repeat is shown with its recognition nucleotide. The protein sequence of a representative repeat is shown with its RVD, HD, in position 12 and 13 highlighted; from the RVD code, HD recognizes C as shown. **B.** Cartoon representations of the crystal structure of PthXo1 wrapped around a 36-bp dsDNA (PDB: 3UGM - Mak *et al.*, 2012) shown in two orientations. The interaction of Repeat 5 (RVD = HD - red sticks) with its target nucleotide (C - cyan stick) is shown on the bottom right.

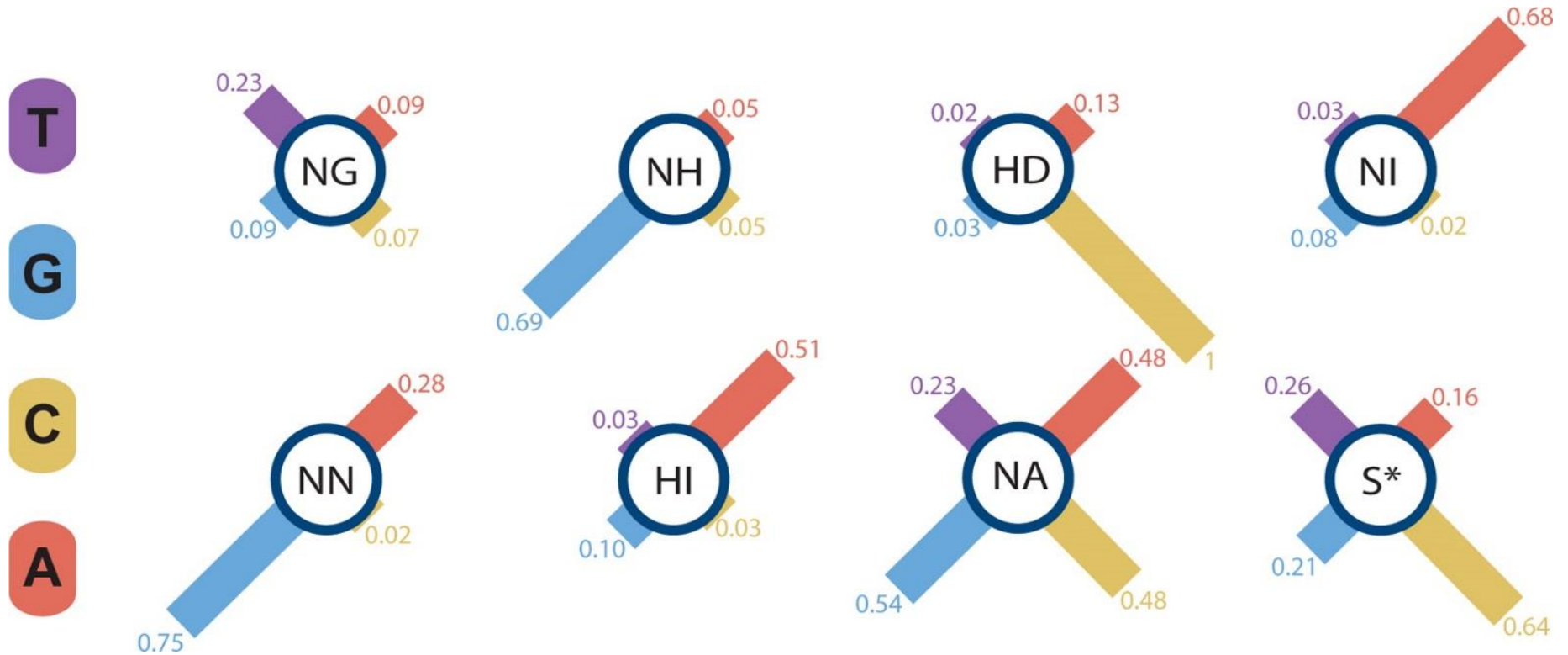


Figure 1.9: Nucleotide preference of commonly used RVDs. RVDs show specific association with one or more nucleotides. RVD affinity for thymine seems to be the lowest across those analysed. However, since NG shows a stronger affinity for T than for other bases, it is usually associated with T in the RVD code for TALE DBD design. Information about the affinities of TALE RVDs for other nucleotides allows the prediction of off-target activities for genome editing work. The information presented in this figure was extrapolated from the work of Cong *et al.*, (2012) by Moore *et al.*, 2014. Cong and colleagues tested a set of 23 TALE proteins with 12.5 RVDs where only RVDs 5 and 6 were substituted for the 23 naturally-occurring RVDs being assayed. A luciferase reporter assay was used to study the transcriptional activity of the 23 TALEs and the data shown here was normalized to the affinity of HD for C (from Moore *et al.*, 2014).

Although the CRD region is responsible for DNA binding, two degenerate repeats (0 and -1) in the N-terminal region are thought to conduct the recognition of the conserved 5'-thymine, with a highly conserved residue within the -1 repeat, Tryptophan 232, making a van der Waals contact with T⁰. Gao *et al.*, (2012) in elucidating the crystal structure of a designer TALE (dTALE2), identified four cryptic repeats in the NTR, designated according to PthXo1 nomenclature repeats 0, -1, -2 and -3. These repeats also form a similar superhelix to the CRD tandem repeats although they do not show nucleotide specificity. They proposed that the NTR repeats bound to DNA in a sequence-independent manner, while identifying residues such as W232, within the NTR that might be essential for TALE binding activity. Single-molecule studies suggest that facilitated diffusion is essential to TALE target location, with the NTR serving as the nucleation site and 'reading domain' for TALE binding while the CRD enhances transitioning into the fully-formed recognition state (Cuculis *et al.*, 2015; Lei *et al.*, 2014). The implications of these NTR interactions for this research work will be discussed in Chapter 4.

TALE proteins fused to nuclease domains such as the FokI endonuclease are called TALENs; they have found significant use in genetic manipulation in mice, pigs, plants and mammalian cells (Christian *et al.*, 2013; Panda *et al.*, 2013; Joung and Sander, 2012). Active fusion proteins of TALE domains with repressors, recombinases, epigenetic modifiers and other effector domains have also been reported (Kubik and Summerer, 2016; Bernstein *et al.*, 2015; Li *et al.*, 2015; Mercer *et al.*, Owens *et al.*, 2013)

The minimal structure of functional TALEs is not yet known, and it does seem that truncations of the N- and C- terminal domains have effects on their binding activities with the former having more deleterious effects on TALE-DNA affinities than the latter (Meckler *et al.*, 2013; Bogdanove *et al.*, 2012; Li *et al.*, 2011; Miller *et al.*, 2011; Mussolino *et al.*, 2011).

In the design of TALENs, two main TALE domain architectures are adopted, the +136/+63 scaffold and the +153/+47 scaffold, with the numbers representing the

number of amino acid codons remaining in the N-terminal and C-terminal domains respectively (Sakuma *et al.*, 2013). The nuclease domains are generally fused via a linker to the C-terminal end of the TALE. The functional spacer length for TALEN architecture varies from 6-40 bp indicating that the flexibility of the TALE region allows the FokI catalytic domains to dimerize across such a wide range of spacer lengths. The spacer region is the central sequence between the target binding sites of the two TALE monomers and is the cleavage site of the nuclease domain. More information on TALE scaffolds and site length requirements is given in Chapter 4.

TALE proteins provide high-fidelity DNA targeting for genome editing work. No evidence of off-target mutagenesis was reported by Li *et al.*, 2011 in targeting an endogenous locus in yeast, and studies in more complex systems using TALENs report similar results, with very minimal off-target activity observed (Guilinger *et al.*, 2014; Mussolino *et al.*, 2011). Careful design and/or selection of target sites can also eliminate off-target activity of TALENs (Kim *et al.*, 2013). TALENs have also been found to be less toxic than ZFNs. This could be because of the length of the conventional TALEN binding site (30-40 bp, excluding the spacer) which theoretically allows a random site to be found only once in 6.2 billion bp (twice the human genome). Bogdanove and colleagues (2011) demonstrated this precise targeting effect of TALE in its ability to discriminate between the highly similar CR2, VP16 and VP64 loci. It is also important to note that different mismatches and their positions have varying effects on TALE activity, with even single mismatches able to disrupt TALE activity entirely (Rinaldi *et al.*, 2017). Other factors that might affect TALE activity are binding site location within the genome and chromatin state; however, the effects of these factors have not been adequately elucidated.

The assembly of designer TALEs has been made quick and efficient using the golden gate cloning system (Cermak *et al.*, 2011). The sequences of TALE repeats (i.e. the DNA TALE-coding sequences) are carefully designed such that restriction endonucleases can cleave the ends of the DNA leaving 4-bp overhangs that enable multiple fragments to be joined in an orderly fashion.

In expanding the TALE toolkit, all possible RVD combinations have been analysed, yielding interesting information for TALE design (Yang *et al.*, 2014). Active TALE variants that have lost the requirement for the initial 5'-thymine have been designed; these provide a potentially limitless targeting opportunity for genome editing (Lamb *et al.*, 2013; Tsuji *et al.*, 2013). Sakuma *et al.*, (2013) also showed that designing TAL effector nucleases, called Platinum TALENs, with additional variations (instead of only in the RVDs) in the repeat regions improves gene disruption activity in frogs and rats.

1.3.1.3 RNA-guided engineered nucleases (RGENs) - CRISPR

'CRISPR' is loosely used to refer to RNA-guided nuclease systems such as the CRISPR/Cas9 and the CRISPR-CPF1 systems. CRISPR systems are responsible for adaptive immunity in some archaea and bacteria. The bacteria generate an array (the CRISPR array) by inserting 'protospacer' sequences captured from invading viruses and genetic elements between direct repeat sequences (Doudna and Charpentier, 2014).

These CRISPR arrays serve as a memory that can be used to tag repeat invaders for destruction by Cas genes. In the CRISPR-Cas9 system, the protospacers are processed into small RNAs (CRISPR RNA or crRNA); this crRNA works with a trans-activating crRNA (tracrRNA or trRNA) to guide the Cas9 protein to the target DNA (Kim and Kim, 2014). Cas9 with its two nuclease domains- a HNH-like domain and a RuvC-like domain, generates a double-strand break at the 20-bp site targeted by the crRNA. This 20-bp sequence is usually flanked at the 3'-end by a 2-6 bp protospacer-associated motif (PAM) sequence. *S. pyogenes* Cas9 requires a 3 bp spacer (NGG) for its activity on the guide-RNA defined target site.

A significant amount of research has gone into simplifying and optimizing CRISPR/Cas9 activity for mammalian genome editing. The crRNA and tracrRNA pair has been simplified into a single construct to generate single guide RNAs (sgRNA) (Jinek *et al.*, 2012). The CRISPR/Cas9 system has been used to induce site-directed

cleavage in endogenous loci in human and mouse cells, triggering repair by NHEJ or HDR (Cong *et al.*, 2013; Mali *et al.*, 2013); for a myriad of genome editing applications in model organisms (Wang *et al.*, 2016); and even for gene correction in human embryos (Ma *et al.*, 2017). Since it does not require multiple cloning steps or selection for design, the CRISPR/Cas9 system is easier to assemble for genome editing in comparison to ZFNs and TALENs. This has made it the prime tool for gene targeting in cell and molecular biology research.

While the wild-type CRISPR/Cas9 system is an effective tool for nuclease-based genome editing, nuclease-deficient Cas9 (CRISPR/dCas9) allows the repurposing of the system for transcriptional repression and activation, epigenetic modifications and even optogenetic applications (Fu *et al.*, 2016; Hilton *et al.*, 2015; Polstein *et al.*, 2015; Shalem *et al.*, 2015).

1.3.2 Genome Editing: HIV-1

Genome editing has been proposed as a unique strategy to inhibit HIV-1 infection and to eradicate the latent viral reservoir by targeting the genome of the host cells (Stone *et al.*, 2013). Two of the approaches taken to do this are discussed below.

1.3.2.1 *Disruption of Host Cellular Receptors*

Host cellular receptors can be genetically disrupted to prevent the attachment and entry of the virus into immune cells. CD4 is the primary attachment receptor for HIV-1; however, it is essential for normal human immune function. Co-receptors C-C chemokine receptor type 5 (CCR5) and C-X-C chemokine receptor type 4 (CXCR4) have been the main targets of such genomic editing strategies. All three nuclease-based technologies- ZFNs, TALENs and RGENs- have been used for this (Xu *et al.*, 2017; Liu *et al.*, 2014; Yao *et al.*, 2012; Perez *et al.*, 2008).

Notably, there are three active clinical trials and three completed trials led by Sangamo Therapeutics exploring ZFN-mediated CCR5 gene disruption to confer HIV-1 resistance to persons infected with HIV (DiGiusto *et al.*, 2016; Tebas *et al.*,

2014). Here, CCR5, the gene for the entry co-receptor for most HIV-1 strains into T-cells, is disrupted to mimic the CCR5 delta-32 mutation found in HIV-resistant individuals. This genetic disruption by ZFNs is carried out *ex-vivo* on hematopoietic stem/progenitor cells (HSPCs) isolated from HIV-1 patients under clinical trials. Patients are then infused with 5-30 billion CCR5-modified HSPCs (SB-728mR-HSPC). This work is built based on previous research that has shown that infused HSPCs have the ability to continuously generate sufficient HIV-1-resistant progeny and thus to rebuild immune systems (Amado *et al.*, 2004; Wang *et al.*, 2009). It will be interesting to see how effective this strategy becomes and how far down the clinical trial pipeline it gets. However, the potential for off-target genome modifications with nuclease-based tools still comes into play here (Wang and Cannon, 2016).

1.3.2.2 *Proviral mutation or excision*

Receptor-targeting approaches are more akin to ART as they do not aim to provide a cure but to halt active disease by avoiding newer infections. The application of genomic editing for targeting latent viral reservoirs could potentially eradicate the infection. This strategy has taken two forms- genomic mutation of HIV promoter DNA sequences or key viral genes, and full genomic excision of the provirus. Directed evolution techniques have been applied in the past to generate programmable nucleases and tyrosine recombinases for these functions (Kaminski *et al.*, 2016; Ebina *et al.*, 2013; Sarkar *et al.*, 2007).

Ebina *et al.* (2013) used the CRISPR/Cas9 system for the genomic mutation of two regions in the LTR of HIV-1 that are essential for viral transcription and replication; the binding site of transcription factor, Nf-kB (nuclear factor kappa-light-chain-enhancer of activated B cells) and the transactivation response (TAR) element (a stem-loop structure that responds to the transactivation protein, Tat) (See Section 3.3.1 for more information on LTR structure). They demonstrated considerable reduction in HIV-1 LTR-driven expression by confronting a pseudo-HIV construct having the full 5' and 3' HIV-1 LTR flanking Tat and GFP genes with the CRISPR/Cas9 components. They showed suppression of GFP expression in

human cell lines, and sequence analysis revealed CRISPR/Cas9-mediated deletions and insertions in the target regions, especially when the TAR element was targeted. It is important to note here that in considering clinical importance, genomic mutation of HIV-1 can be a self-limiting strategy. As mentioned earlier, HIV-1 has an error-prone reverse transcriptase: the virus can tolerate significant mutations. For increased effectiveness, multiple essential HIV genes can be disrupted simultaneously.

In 2015, Ebina and co-workers reported the targeting of the TAR element using TALENs. Their design here focused on proviral excision, not just genomic mutation. They observed proviral excision in more than 80% of cell lines. The TALENs also offered higher precision with no off-target activities detected. Others who have used the CRISPR/Cas9 system for HIV-1 provirus disruption and/or excision include Yin *et al.*, (2017), Huang and Nair, (2017), Kaminski *et al.*, (2016) and Dampier *et al.*, (2014).

1.3.3 Genome Editing: Limitations of nuclease-based systems

The reliance of nuclease-based genome editing strategies on host cell repair mechanisms (NHEJ and HDR) leads to wide variability in therapeutic outcomes. Off-target activities of nuclease catalytic domains as well as fatal indel mutations from these repair mechanisms could have debilitating effects on the host cells. The different activities of these repair mechanisms in different cell types and growth stages also place limitations on the clinical applications of nuclease-based genome editing.

A recent publication (Wang *et al.*, 2016) demonstrates that although CRISPR/Cas9 HIV-1 targeting inhibits viral replication, it also results in increased development of viral resistance to the targeting mechanism. This is because some mutations generated by indels confer resistance to the virus. A commentary by Liang *et al.*, (2016) on the limitations of the CRISPR-Cas9 system for tackling HIV infections also provides a good outlook on the subject.

Despite the fidelity of the DNA-binding proteins used for targeting HIV-1 (CRISPR RNA guides, TAL effector proteins or Zinc finger proteins), the drawbacks of nuclease-based systems currently prevent clinical therapeutic applications of genome editing for HIV. For precise human gene therapy, nuclease domains might be replaced with other effector domains, such as recombinases, that have the inherent ability to cut and re-join DNA.

1.4 Site-Specific Recombination

Site specific recombination involves the cleavage and re-joining of DNA at two specific sites. There is no requirement for extensive sequence homology at the sites, or ATP or other high-energy co-factors (Grindley *et al.*, 2006). The process is directed by enzymes referred to as site-specific recombinases. These enzymes are encoded by bacteria, archaea, bacteriophages and eukaryotes. They are often important for the mobility of mobile elements such as in the resolution of transpositional cointegrates or the integration of bacteriophage DNA into the genome.

Recombinases identify and bind to unique recognition sites, then catalyse the recombination reaction. The DNA rearrangement actions mediated by these enzymes can be integration, excision or inversion (Fig. 1.10). If the two sites are on two different molecules with either or both molecules being circular, recombination will result in the fusion of both molecules into one. If the two sites are on the same molecule and in direct repeat, the recombinase will catalyse excision (resolution) of a DNA circle and yield two molecules. If the two sites on a single DNA molecule are in inverted repeat, inversion of the sequence between the sites relative to the remaining sequence will occur.

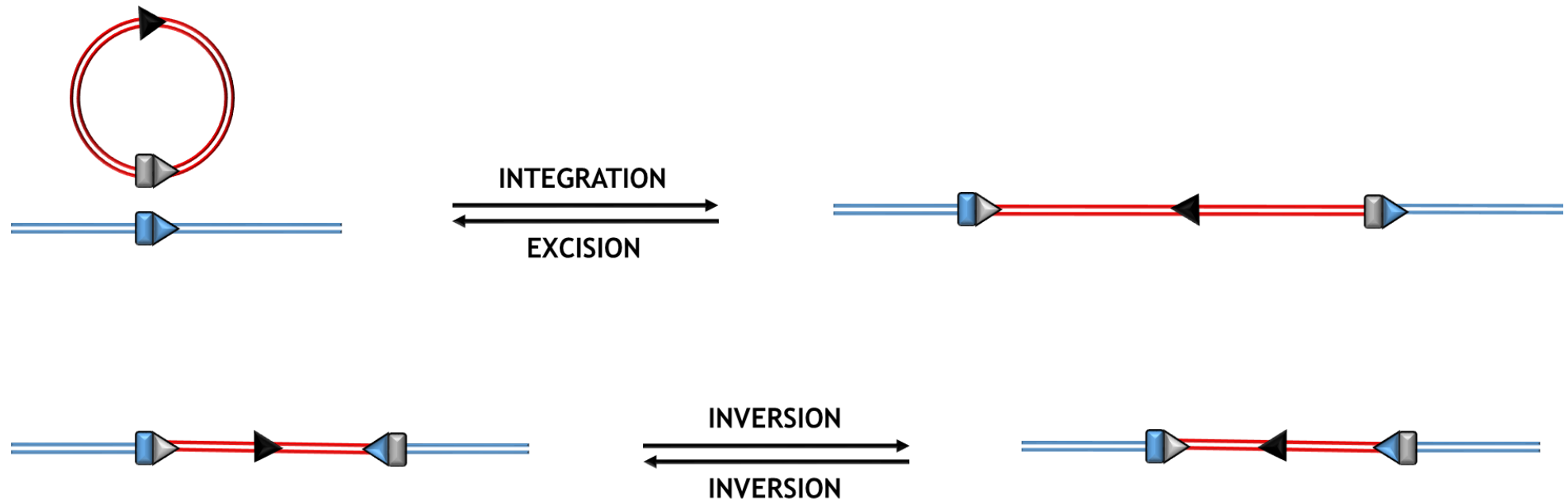


Figure 1.10: Recombinases promote several modes of DNA rearrangements. This depends on the orientation of both sites. A. A circular DNA molecule becomes integrated into the linear DNA and the reverse reaction occurs to excise out the DNA when the sites are in direct repeat, reforming the original circular conformation. B. Recombination between inverted repeat sites results in the inversion of the DNA orientation between the sites.

Two unrelated families of site specific recombinases (SSRs) have been identified; the tyrosine recombinase family and the serine recombinase family, both named for the conserved nucleophilic residue responsible for DNA attack and protein-DNA complex formation during the recombination reaction (Olorunniji *et al.*, 2016). Although both families of recombinases carry out the same biological functions, they have mechanistic and structural differences.

1.5 Tyrosine Recombinases

Tyrosine recombinases are found primarily in prokaryotes and bacteriophages where their functions range from DNA integration and excision to plasmid number control. They have also been identified in some eukaryotes and archaea (Jayaram *et al.*, 2015).

X-ray structures of different tyrosine recombinases with their target DNA show a highly similar tetrameric complex where each of the monomers interacts with both its target DNA and a neighbouring monomer (Meinke *et al.*, 2016) (Fig. 1.11). Only two of the four monomers are in an active catalytic conformation at any one time. As such, the recombination mechanism of tyrosine recombinases is step-wise. Cleavage occurs with single-strand breaks at cleavage sites which are 6-8 bp apart (Meinke *et al.*, 2016). The nucleophilic tyrosine residue of active monomers attacks the scissile phosphate at its target site resulting in 3' phospho-tyrosine and 5'hydroxyl ends. Each free 5' hydroxyl end then attacks the neighbouring phospho-tyrosine linkage resulting in a Holliday-junction intermediate. Conformational changes occur that activate the other two monomers. Strand cleavage and exchange occurs as before, completing the recombination process.

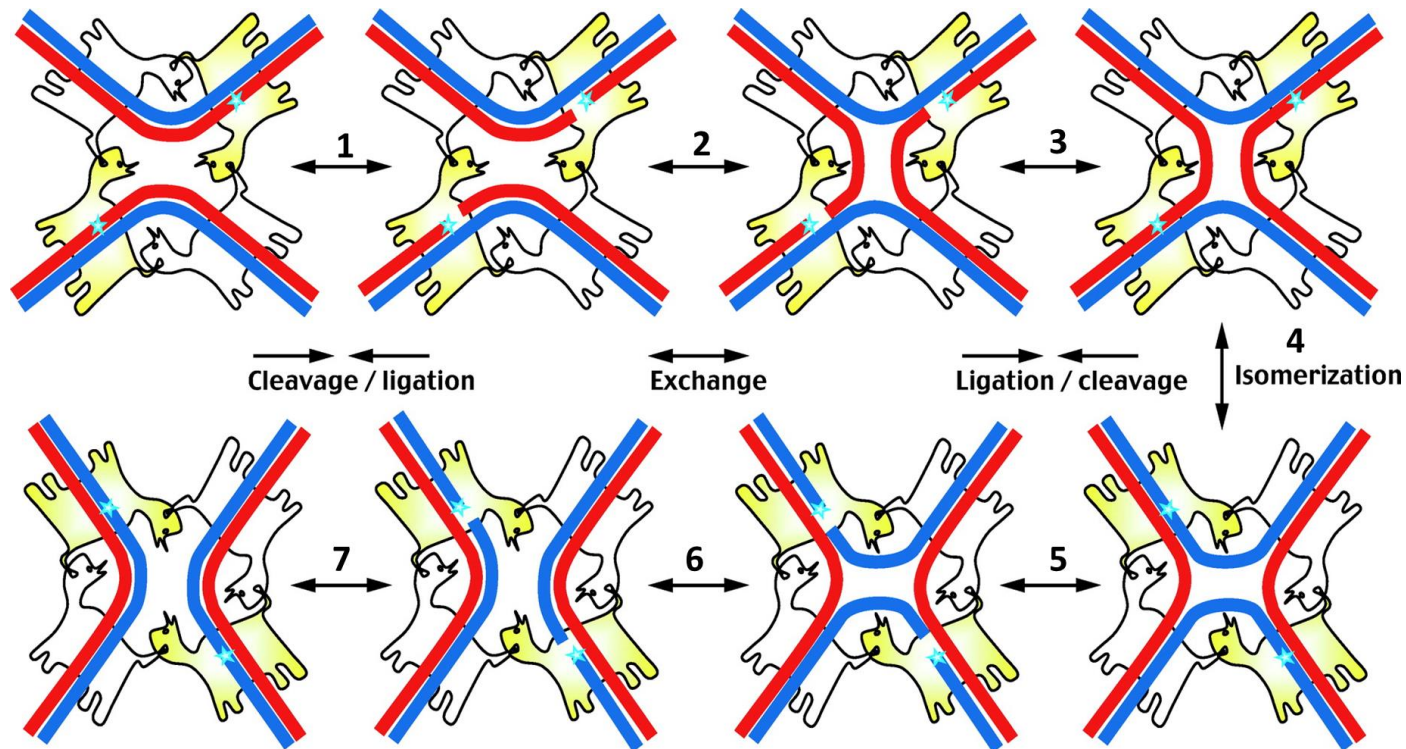


Figure 1.11: Cartoon depicting the mechanism of strand exchange by tyrosine recombinases. A synaptic complex of four recombinase monomers bound to two DNA duplexes is shown (1). Monomers in the active cleaving catalytic conformation are coloured yellow, with the blue stars indicating the catalytic tyrosine nucleophile. One strand of each duplex is first cleaved (2) and then exchanged (3), forming an Holliday junction intermediate (4). The catalytic activation of the second pair of monomers by isomerization of the junction leads to a second stage of cleavage (5), strand exchange (6) and ligation (7), yielding recombinant products (originally from Grindley *et al.*, 2006)

The N- and C- terminal domains are both important for DNA-binding interactions. The C-terminal domain also contains all the active catalytic residues. This lack of functional modularity and the complex protein-protein interactions important for transition between active and inactive conformations limit their reprogramming to target different DNA sequences.

Tyrosine recombinases such as Cre recombinase, Flp recombinase and the λ integrase have found significant use in biomedical and biotechnological research and in clinical applications. Directed evolution approaches have also been applied to retarget the Cre recombinase for HIV-1 proviral excision.

1.5.1 Tre and Brec1: reprogrammed tyrosine recombinases for HIV-1 excision

Cre recombinase is one of the most studied of the tyrosine recombinases. It is a 343-amino acid tyrosine recombinase from Bacteriophage P1 whose natural function is in DNA cyclization and resolution after replication (Van Duyne, 2015). It carries out excision, integration and inversion of DNA between two identical 34-bp recognition sequences called '*loxP*' (locus of X(cross)-over in P1) sites. The recombination outcome is based on the relative orientation and location of the two sites (Fig. 1.10). Since its discovery in 1981, the Cre-*loxP* system has become a significant genetic tool for manipulating eukaryotic genomes *in vivo* and *in vitro*.

In 2007, Sarkar and colleagues of the Hauber and Buchholz research group evolved a Cre recombinase variant to target HIV-1. They chose a 34-bp target sequence '*loxLTR*' that is similar to the cognate *loxP* site. *loxLTR* comes from the LTR of a rare HIV-1 isolate TZB0003. However, TZB0003 has an atypical LTR sequence that is divergent from most HIV-1 strains and represents less than 1% of sequenced primary isolates. The mutant Cre, called Tre recombinase was identified after 126 rounds of a directed evolution approach called substrate-linked protein evolution (SLiPE) (Sarkar *et al.*, 2007). SLiPE is a directed evolution approach that combines

the coding sequence of the recombinase being engineered with its target recognition DNA sequence on the same plasmid; allowing the screening of huge libraries for rare mutants (Buchholz and Stewart, 2001). Since then, the ability of this protein to target HIV has been demonstrated using different approaches. Tre has 19 mutations relative to wild-type Cre recombinase. The recombination efficiency of Tre in mammalian cells was determined by measuring luciferase activity. Conditional expression of Tre recombinase in HIV-1 infected humanized mice cells from a self-inactivation vector (SIN) resulted in proviral excision with significant reduction in viremia and no induction of Tre-mediated cytopathic effects (Hauber *et al.*, 2013). The crystal structure of a catalytically inactive Tre recombinase variant and its mutational analysis have also recently been carried out (Meinke *et al.*, 2017) (Fig. 1.12). This revealed that 9 of the 19 mutations in Tre recombinase interact with the DNA (3 with direct sequence-specific readout and 6 interacting with the DNA backbone). The remaining mutations are thought to be important for conformational adaptability of the protein to its new target site. 17 of the 19 mutations were shown to be important for Tre recombinase activity. Still, the *loxLTR* target site is not clinically significant for HIV-1 genomic editing.

To develop a clinically relevant tool, the research group scanned through the LANL HIV database and identified a new target site that is both highly conserved in many HIV-1 isolates and slightly similar to the *loxP* site. The new target site, *loxBTR* is a 34-bp highly conserved sequence in the R region of the LTR present in about 90% of the HIV-1 isolates of subtypes A, B and C (Karpinski *et al.*, 2016). These subtypes represent 72% of confirmed and sequenced global HIV-1 cases; this is definitely a clinically relevant target. *loxBTR* has 11 nucleotides similar to the native 34-bp *loxP* site.

They then proceeded to carry out 145 substrate-linked protein evolution rounds on Cre recombinase to generate new mutants. The protein evolution rounds also contained 18 rounds to exclude recombination on similar target sequences and thus improve specificity. Of over 100 clones at round 145, the Brec1 mutant was selected as it recombines *loxBTR* and not *loxP* or target sequences of related

proteins. It also tolerates point mutations, recombining point-mutated *loxBTR* sequences with the four most common mutations in the Brec1 target region of primary HIV-1 isolates. They infer that this will allow the protein to tolerate some viral resistance. Brec1 activity has been demonstrated in bacteria, mammalian cells and in mice humanized with patient-derived cells with no reported observable cytotoxicity effects (Karpinski *et al.*, 2016). This is an impressive step forward in the generation of a tool capable of specifically excising HIV-1 *in vivo* in infected patients.

However, there are limitations to the use of Tre, Brec1 and other tyrosine recombinase-sourced proviral excision tools. Although more structural and mechanistic information is being generated about the target site recognition and catalytic requirements of Cre recombinase, the bases for these are still not quite clear. The lack of functional modularity of tyrosine recombinases has been mentioned earlier. Since both the N-terminal and C-terminal domains are required for target site recognition and the mediation of catalysis, a new protein would have to be engineered each time the virus develops intolerable mutations at the recombinase target site. Karpinski *et al.*, (2016) predict the development time for a new enzyme to be two years; this is a significantly long time for a highly mutagenic pathogen like HIV-1.

The structural and functional modularity of serine recombinases makes them easier to engineer than the tyrosine recombinases.

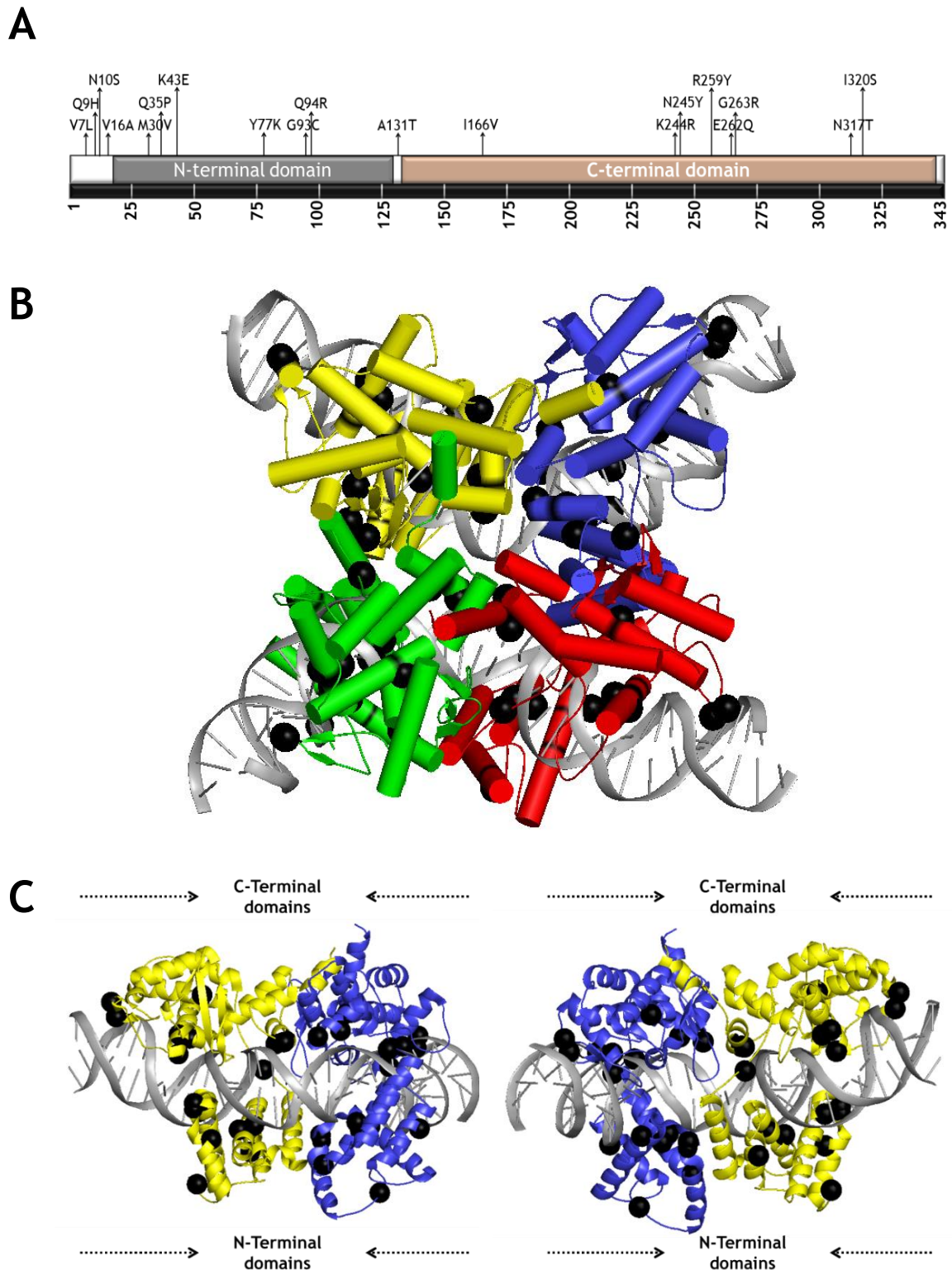


Figure 1.12: Mutations in Tre recombinase relative to Cre recombinase. A. Location of mutations across the protein **B.** Cartoon representation of crystal structure of Tre recombinase (PDB:5U91- Meinke *et al.*, 2017) **C.** Dimers shown from two perspectives. The four monomers are shown in yellow, blue, red and green barrels or ribbons with the mutated residues shown as black spheres and the DNA is shown as grey ribbons. The 19 mutations in Tre recombinase are spread across the N- and C- terminal domains.

1.6 Serine Recombinases

Previously referred to as the invertase/resolvase family, serine recombinases are a heterogeneous family of recombinases consisting of invertases, transposon resolvases, phage integrases and others. The proposed mechanism of action of serine recombinases is illustrated in Fig. 1.13. Unlike tyrosine recombinase there is no formation of an intermediate Holliday junction. They also differ from tyrosine recombinases in that their two domains are functionally and structurally modular.

For gamma-delta ($\gamma\delta$) and Tn3 resolvase, two transposon-derived small serine recombinases, the mechanism involves the formation of a tetramer between two aligned crossover sites. Each recombination “crossover” site is first bound by two recombinase subunits; two of these recombinase-bound crossover sites then come together to form a synaptic complex consisting of four recombinase subunits and two DNA crossover sites. This proceeds to concerted cleavage of the two sites with a staggered 2-bp overlap yielding four half-sites with a 5'-phosphoserine recombinase-DNA complex and a 3'-hydroxyl end on each half-site. Cleavage occurs as a result of the attack on the scissile phosphodiester bond of the DNA backbone by the nucleophilic serine residue. After cleavage, strand exchange involving a 180° subunit rotation followed by re-ligation of the half-sites occurs (Fig. 1.13).

Wild-type serine recombinases are generally very efficient and specific, with their activities tightly regulated. They have strict recognition sequence specificity, spacing and topology requirements, and some require accessory factors (Fig. 1.14). These accessory factors, usually additional recognition DNA sequences (called accessory sites) or proteins, can be required for binding of other recombinase subunits or formation of complex structural intermediates (Stark, 2014). Topology requirements include negative supercoiling of the DNA substrate and specific relative orientation of the sites.

The main recombinase utilized in this research work is the 20-kDa Tn3 resolvase. Very similar to the well-researched and characterized $\gamma\delta$ resolvase, its natural function is co-integrate resolution during transposition by the Tn3 transposon. The Tn3 transposon is a mobile genetic element that confers β -lactam antibiotic resistance to *E. coli* and other gram-negative bacteria through horizontal gene transfer. It encodes three genes, Tn3 resolvase (*tnpR*), Tn3 transposase (*tnpA*) and β -lactamase (*Bla*) and is flanked at both ends by a 38-bp sequence in an inverted orientation.

During replicative transposition, a 4957-bp Tn3 transposon and its copy is inserted (in direct repeat) into the host plasmid DNA with the aid of the transposase (Grindley, 2006; Shapiro, 1979) (Fig. 1.14). This results in the formation of an intermediate called the transpositional cointegrate. Tn3 resolvase then mediates site-specific recombination at sites called '*res*' present on each copy of the transposon, resulting in the resolution of the co-integrate into two circular molecules.

Tn3 resolvase (or $\gamma\delta$ resolvase) recombines two *res* sites of 114-bp. Each *res* site contains three resolvase binding sites; site I (28 bp), at the centre of which is the crossover point where strand cleavage and re-joining occurs; and sites II and III, of 34-bp and 25-bp respectively that serve as accessory sites essential for the precise assembly of the synaptic complex necessary for recombination catalysis (Fig. 1.14). The final product of Tn3 resolvase activity is a 2-noded catenane, two topologically interlocked DNA circles, which can then be separated by a type II topoisomerase (Stark *et al.*, 1989).

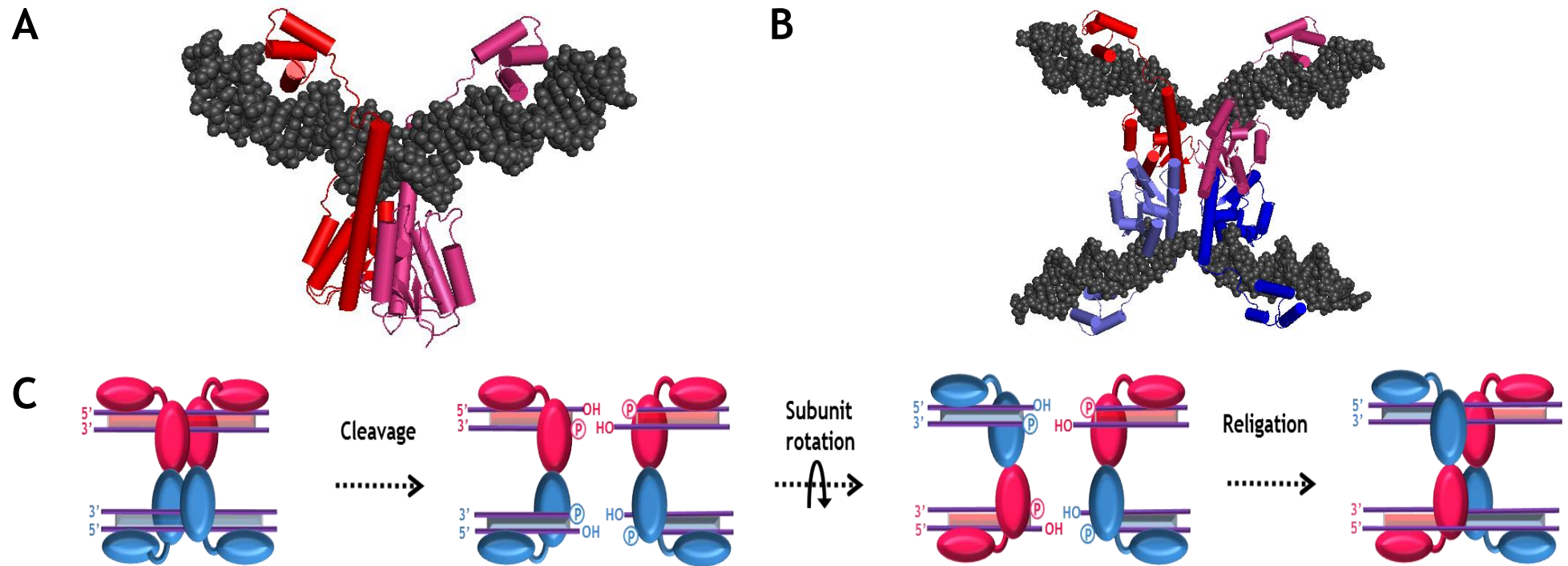


Figure 1.13: Site-specific recombination - structure and mechanism. A. 1GDT, the crystal structure of wild-type $\gamma\delta$ resolvase (proposed to be similar for Tn3 resolvase) dimer bound to 34-bp *res* site 1 (PDB:1GDT- Yang and Steiz, 1995). B. 1ZR4, crystal structure of activated mutant of $\gamma\delta$ resolvase tetramer bound to two cleaved *res* site I. *res* site I DNA and backbone are shown as blue spheres in space-filling model, resolvase monomers are shown as red, pink, light blue and dark blue rods and strands (PDB: 1ZR4- Li *et al.*, 2005). C. Simplified proposed mechanism of action of serine recombinases. Pink and blue spheres represent resolvase monomers. A four-strand concerted break and re-join model is proposed (Adapted from Olorunniji and Stark, 2010).

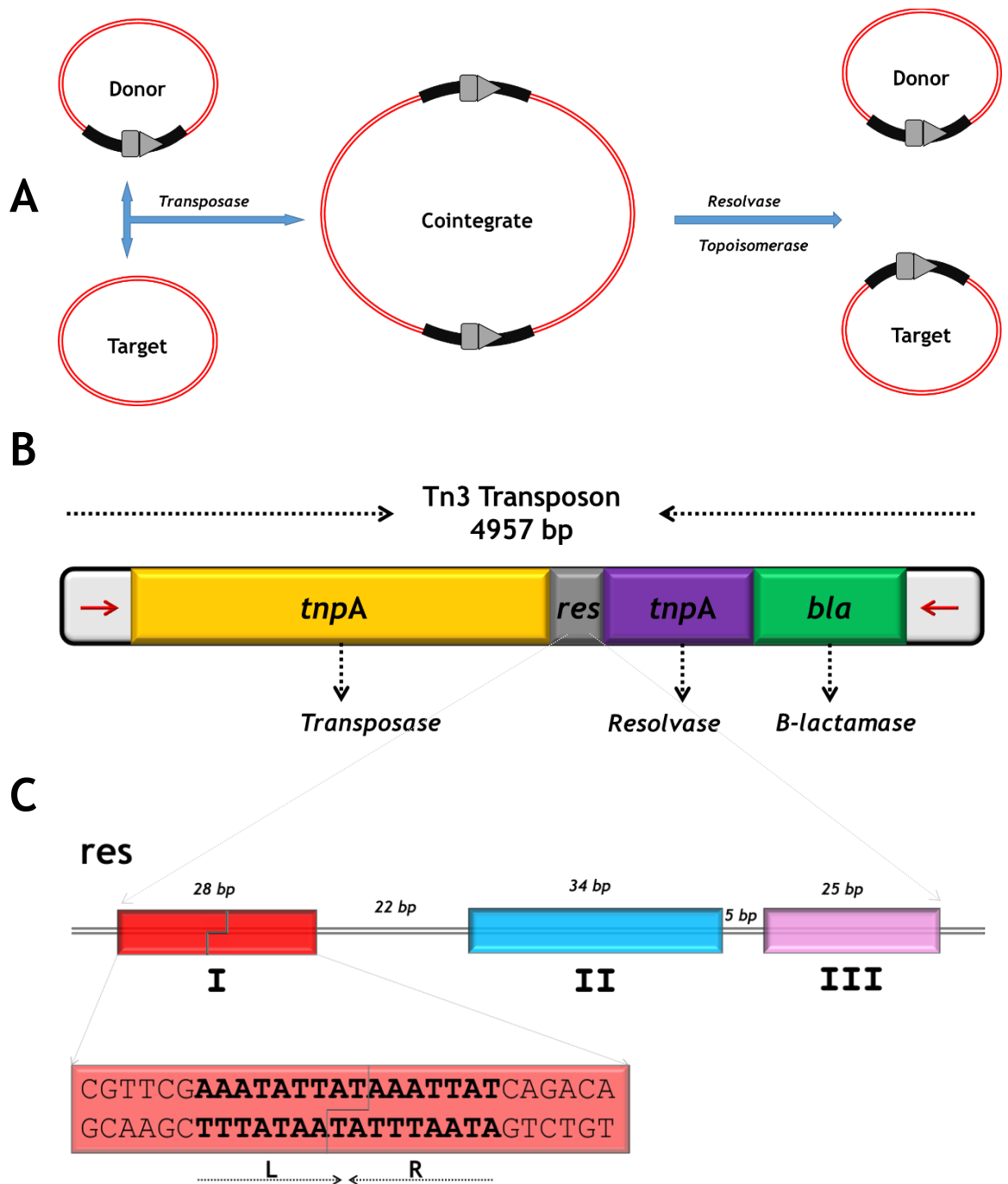


Figure 1.14: Wild-type Tn3 (and $\gamma\delta$) resolvase has strict substrate specificity requirements. **A.** Transposon resolvases are responsible for separating two copies of transposons in a cointegrate intermediate during replicative transposition. This results in a two-noded catenane which goes on to be separated into single circles by the action of topoisomerases. **B.** Tn3 transposon encodes three genes- resolvase (*tnpR*), transposase (*tnpA*) and β -lactamase (*Bla*)- and is flanked at both ends by a 38-bp sequence (recognised by the transposase enzyme) in an inverted orientation. It also carries a 114-bp '*res*' site, the recognition sequence of the resolvase. **C.** The *res* sites of Tn3 ($\gamma\delta$) resolvases have strict nucleotide lengths and spacing requirements (22 bp between *res* site I and II and 5 bp between II and III). The staggered cleavage site on site I is depicted.

1.7 The structure of Tn3 resolvase

Like other small serine recombinases, Tn3 resolvase has two functional domains- an N-terminal catalytic domain and a C-terminal DNA-binding domain. The N-terminal domain contains the active site residues and the nucleophilic Ser 10 responsible for strand cleavage, and is connected to the C-terminal domain by an arm region (Yang and Steiz, 1995). The C-terminal domain is largely responsible for sequence-specific DNA recognition. This functional domain organization is an important feature that this research work is harnessing in the reprogramming of Tn3 resolvase for genome editing.

Most of what is known about the structure of Tn3 resolvase comes from studies on $\gamma\delta$ resolvase. The crystal structures of both the dimeric and tetrameric forms of $\gamma\delta$ resolvase have been extensively studied and $\gamma\delta$ resolvase shares almost 80% sequence similarity to Tn3 resolvase (Li *et al.*, 2005; Yang and Steiz, 1995). It must also be noted here that the structural description of $\gamma\delta$ resolvase is expressed differently by different authors, especially when it comes to the E-helix and 'arm' region of the protein (Fig. 1.15). I will attempt here to provide a simplistic overview.

The N-terminal catalytic domain of $\gamma\delta$ resolvase consists of a five-stranded mixed β - sheet surrounded by four α helices (residues 1-100) and an extended arm region made up of the E-helix (residues 101-137). A short linker (residues 138 - 148) links the E-helix to the C-terminal DNA binding domain, a three-helix bundle (residues 149 - 183) (Fig. 1.15).

Residues 1 to 120 on the N-terminal domain have the same globular structure in the presence and absence of DNA (Sanderson *et al.*, 1990; Rice and Steiz, 1994; Yang and Steiz, 1995). This region contains the nucleophilic S10 and other highly conserved residues essential for active site formation and/or for catalysis. It is also responsible for higher-order interactions between dimers at the synaptic interface. The catalytic domains undergo a structural transformation on synapsis

of two dimers. This is due to the movements of residues 1-100 and the E-helices, creating a flat hydrophobic interface between 'halves' of the structure that allows subunit rotation. A third interface, the 2-3' interface, is formed between catalytic subunits at site I and regulatory subunits, and is thought to induce activation of conformational changes to regulate catalysis through these contacts (Fig. 1.16) (Sarkis *et al.*, 2001).

The second half of the E-helix at the C-terminal end (residues 121 - 137) wraps around the minor groove of the central 16-bp of site I. It seems to be unstructured in the absence of DNA; this character, analogous to the formation of the basic leucine zipper, is attributed to DNA binding. Close packing interactions of the protein and DNA at R125, T126, R130 and F140 and sequence-specific interactions in the minor groove with G141 and R142 have been reported. All these indicate that in addition to direct recognition by the C-terminal domain of the subunit, the N-terminal catalytic domain of wild-type $\gamma\delta$ (and Tn3) resolvase has intrinsic DNA recognition properties (Rice, 2015).

The C-terminal domain interacts with the major groove of the DNA at the outer boundaries of the crossover site. Two of the three helices here form a canonical helix-turn-helix motif responsible for sequence-specific DNA recognition. At least eight salt bridges and hydrogen bonds attach the C-terminal domain into the major groove of the DNA at A161, S162, A171, S162, T174, Y176, K177 and R148 (the final residue in the linker strand). Base-specific interactions to four base pairs at the outer parts of site I occur with R172, S173 and Y176 (Yang and Steiz, 1995).

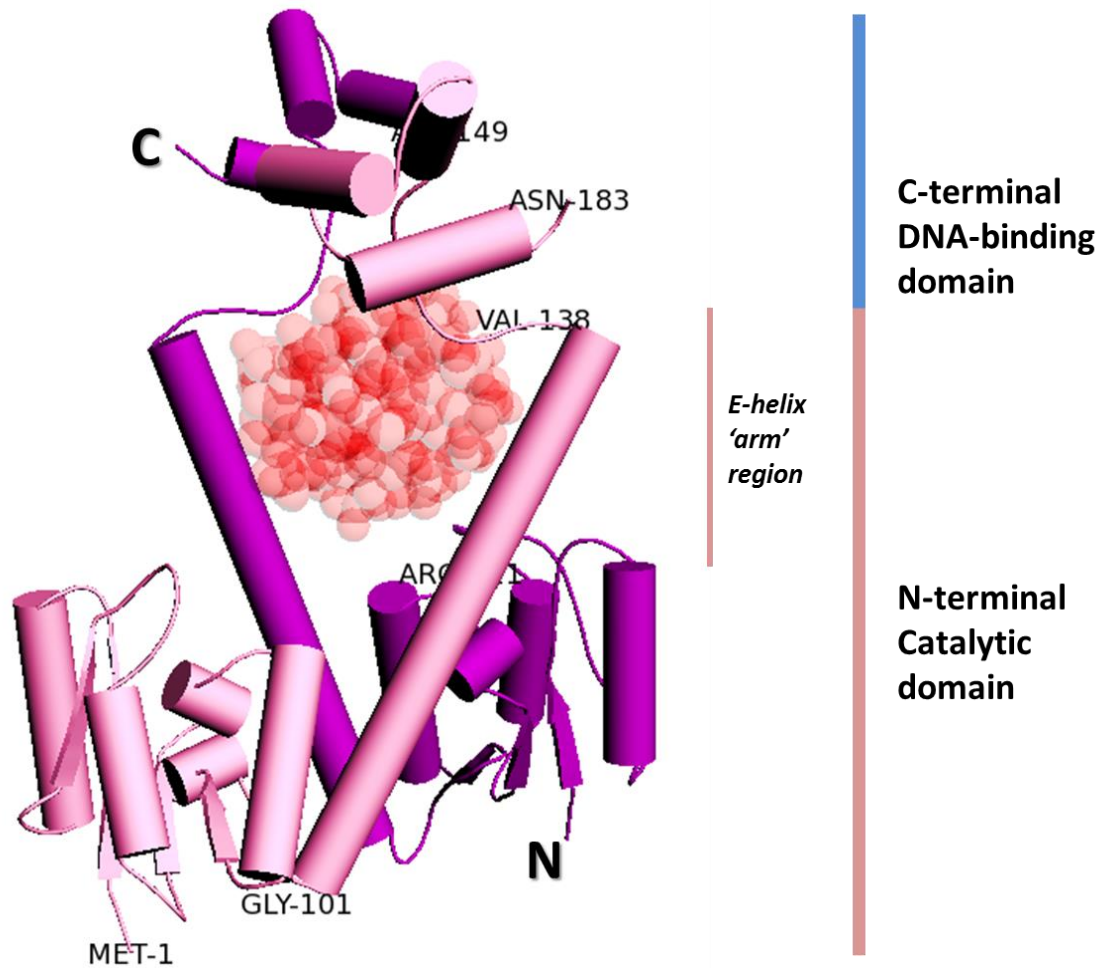


Figure 1.15: The structure of $\gamma\delta$ (and Tn3) resolvase. The modular structure of the $\gamma\delta$ resolvase is illustrated here on a cartoon rendering of the 1GDT crystal structure (Yang and Steiz, 1995). Two monomers of resolvase (pink and purple cylindrical helices) associated in a dimer-complex with DNA (transparent red spheres) are shown with some residues labelled.

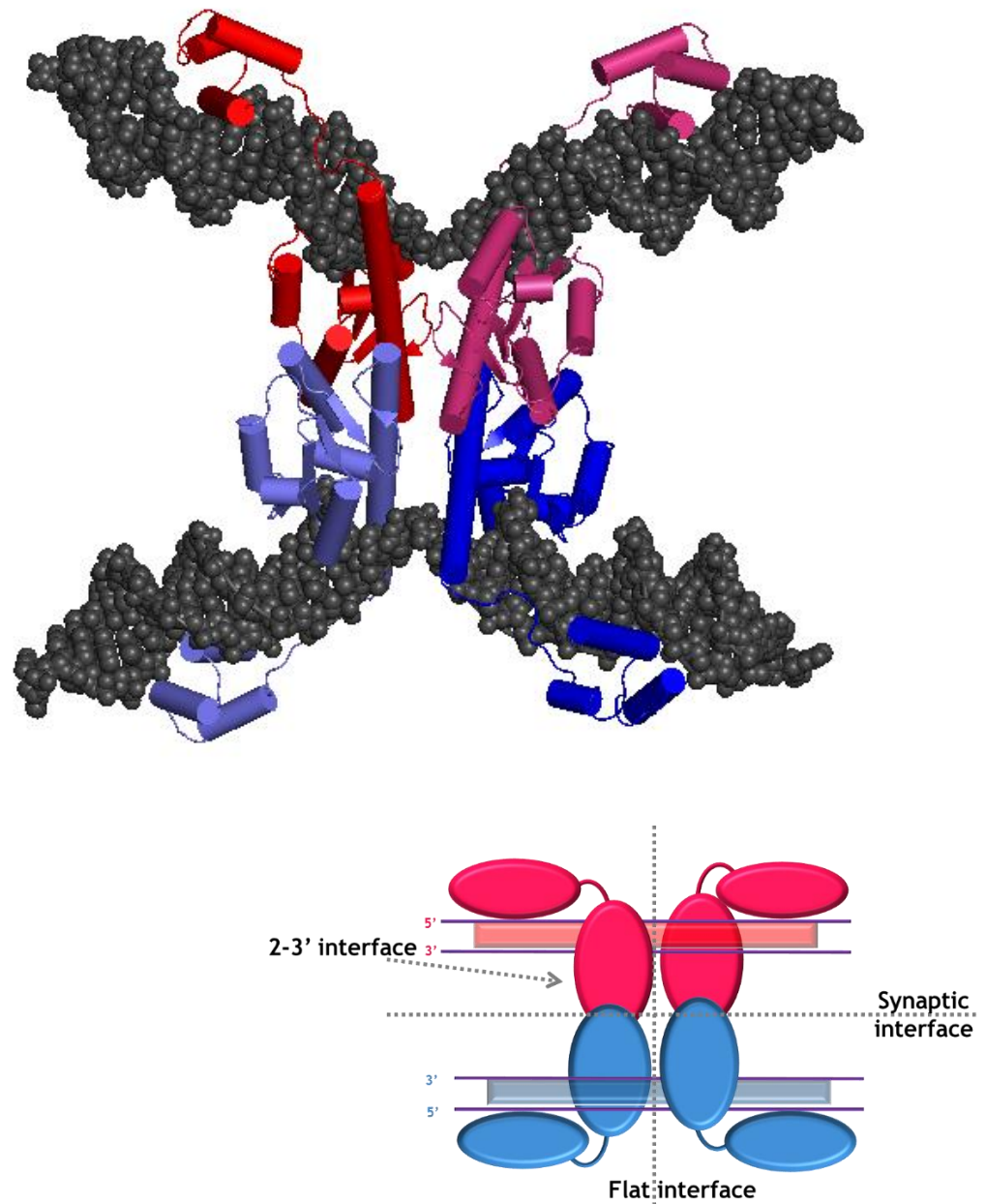


Figure 1.16: Interfaces of $\gamma\delta$ -resolvase. The structure at the top shows the post-cleavage synaptic tetramer of $\gamma\delta$ -resolvase as described in Figure 1.13B (PDB: 1ZR4- Li *et al.*, 2005). A cartoon representation at the bottom highlights the interfaces mentioned in the text. Resolvase monomers are shown in red and blue and the DNA in grey.

1.8 Hyperactive (Deregulated) mutants of serine recombinases

To begin considering the use of site-specific recombinases in genome editing, the severe constraints (requirement for DNA supercoiling, requirement for accessory sites, etc.) required for Tn3 resolvase activity had to be removed or reduced. Several strategies have been applied to obtain ‘deregulated’ or ‘hyperactive’ mutants that have no requirement for accessory sites (Arnold *et al.*, 1999; Burke *et al.*, 2004; Olorunniji *et al.*, 2008) (Fig. 1.17). These mutants have also lost dependence on supercoiling and specific relationship of the two recombining sites (i.e. requirement for direct repeat of sites).

Deregulating mutations are predicted to stabilize or destabilize interactions, intermediates and/or interfaces during recombination (Arnold *et al.*, 1999; Burke *et al.*, 2004; Olorunniji *et al.*, 2008) (Fig. 1.17). These mutants were identified by screening for mutants that catalysed recombination on modified Tn3 *res* substrate plasmids with one or the two target sites having no accessory sites using a colorimetric selection assay. Unlike wild-type proteins, deregulated mutants can catalyse recombination on two site I_s in inverted or direct repeat, on one or two molecules, and also on nicked and linear substrates.

One of the most active deregulated mutants of Tn3 resolvase (also the starting mutant in this work) is the Tn3 NM mutant: it has six mutations relative to the wild-type- R2A E56K G101S D102Y M103I Q105L) (Olorunniji *et al.*, 2008). The R2A and E56K mutations are thought to inhibit 2-3' interactions of resolvase, allowing the formation of site I synapse in the absence of accessory sites. The other mutations might affect the behaviour of resolvase at site I; forming extensive hydrophobic interactions, destabilizing the dimer interface, or changing the property of a ‘hinge’ in the E-helix thought necessary for resolvase activity (Burke *et al.*, 2004; Olorunniji *et al.*, 2008). The next step taken to switch sequence specificity of resolvases was done by generating chimaeric recombinases.

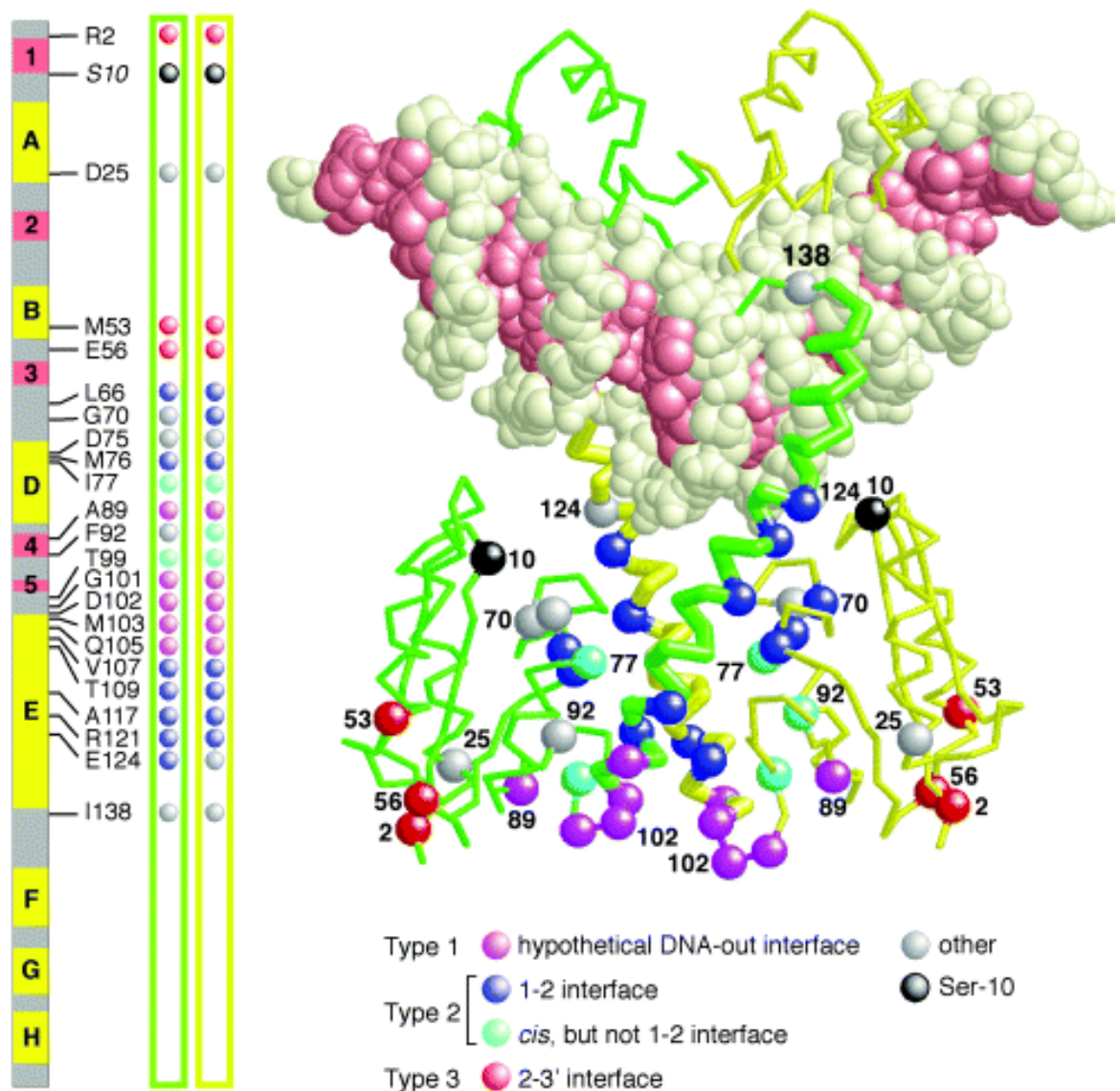


Figure 1.17: Activating mutations of Tn3 resolvase. Mutations of Tn3 resolvase that enhance hyperactivity (recombination without requirement for accessory sites at one or both target sequence(s)) are mapped on the 1GDT crystal structure (Yang & Steiz, 1995). The location of the residues at the interfaces on each monomer gives an indication of mechanism of ‘loss of regulation’. (Burke *et al.*, 2004)

1.9 Chimaeric Recombinases

Chimaeric recombinases (CRs) exploit the modular domain architecture of small serine SSRs to mediate site-specific recombination on non-cognate sites. As the C-terminal domain has no catalytic activity and is only responsible for recognizing and targeting the recombinase, it can be swapped with a different DNA-binding protein domain to target the catalytic N-terminal domain to a modified site. The first approach to generating CRs involved the swapping of C-terminal residues or domains of related serine recombinases such as Tn3 and Tn21 to switch their sequence specificity (Schneider *et al.*, 2000; Avila *et al.*, 1990). While these approaches provided significant mechanistic information, only minimal specificity shift was observed.

A different type of CRs was reported by Akopian *et al.*, 2003, where the C-terminal DNA-binding domain of Tn3 resolvase was replaced with Zif268, the well-characterized zinc finger protein DBD of the mouse transcription factor (EGR-1). Other programmable DNA-binding modules such as TAL effector proteins and the CRISPR-Cas9 have been explored for CR design as well (Chaikind *et al.*, 2016; Mercer *et al.*, 2012).

While the DNA-binding domain is swapped to guide the enzyme to non-cognate target sites, the N-terminal catalytic domain has also been engineered to mediate novel or altered specificity (Proudfoot *et al.*, 2011). The N-terminal domain has been subjected to a range of targeted and random genetic engineering techniques and several residues have been identified that are important for shifting the fixed specificity of SSRs such as Tn3 resolvase. (Gaj *et al.*, 2014; Proudfoot *et al.*, 2011; Sirk *et al.*, 2014). See Chapter 3 for a detailed outline of the strategies used in engineering the N-terminal catalytic domain. These retargeting studies have also suggested that mutating some residues within the E-helix region can confer novel or broadened catalytic specificity. Combining these approaches can hypothetically provide altered-specificity CRs capable of mediating site-specific recombination on any genomic sequence of interest.

1.9.1 Zinc Finger Recombinases

The first zinc finger recombinase was reported by Akopian *et al.*, 2003. The design here involved the replacement of the C-terminal domain of Tn3 NM resolvase with a well characterized Cys2-His2 Zif268 zinc-finger domain, joining the two domains with a short linker. The extreme bases of site I were also swapped with the 9-bp recognition sequence of Zif268, creating a modified site I which was called the Z-site (Fig. 1.18). The chimeric protein, called a zinc finger recombinase, ZFR (originally Z-resolvase), was shown to catalyse efficient recombination reactions on the Z-sites *in vivo* and *in vitro* (Akopian *et al.*, 2003; Prorocic *et al.*, 2011). A different research group has also reported several designs of ZFRs using two-finger and four-finger zinc finger proteins as well as with a different catalytic domain, Gin recombinase (Gersbach *et al.*, 2010; Gordley *et al.*, 2007).

ZFRs consist of two to four zinc finger motifs attached to the recombinase catalytic domain by a short linker (Fig. 1.18). The substrate site has a similar architecture to *res* site I acted upon by deregulated resolvase mutants and consists of a central 22-bp recombinase recognition site flanked at both ends by 9- to 12-bp zinc finger recognition sequence in an inverted orientation. Usually, only the core 16 bp of *res* site I is retained in the design.

Engineering approaches have been applied to extend substrate specificity between recombinases, such as mutating a Tn3 resolvase-based ZFR to recognize Sin resolvase Z-sites (Proudfoot *et al.*, Unpublished data). Sin resolvase is a serine recombinase from *Staphylococcus aureus* and carries out excision between directly repeated sites (Rowland *et al.*, 2002). Its *res* site I target sequence is GC-rich in contrast to that of Tn3 resolvase. ZFRs have also been proposed for use in functional biomedical research such as in bacteria-based bovine b-casein gene intron excision and ZFR-mediated human genomic DNA integration (Proudfoot *et al.*, 2011; Gaj *et al.*, 2013).

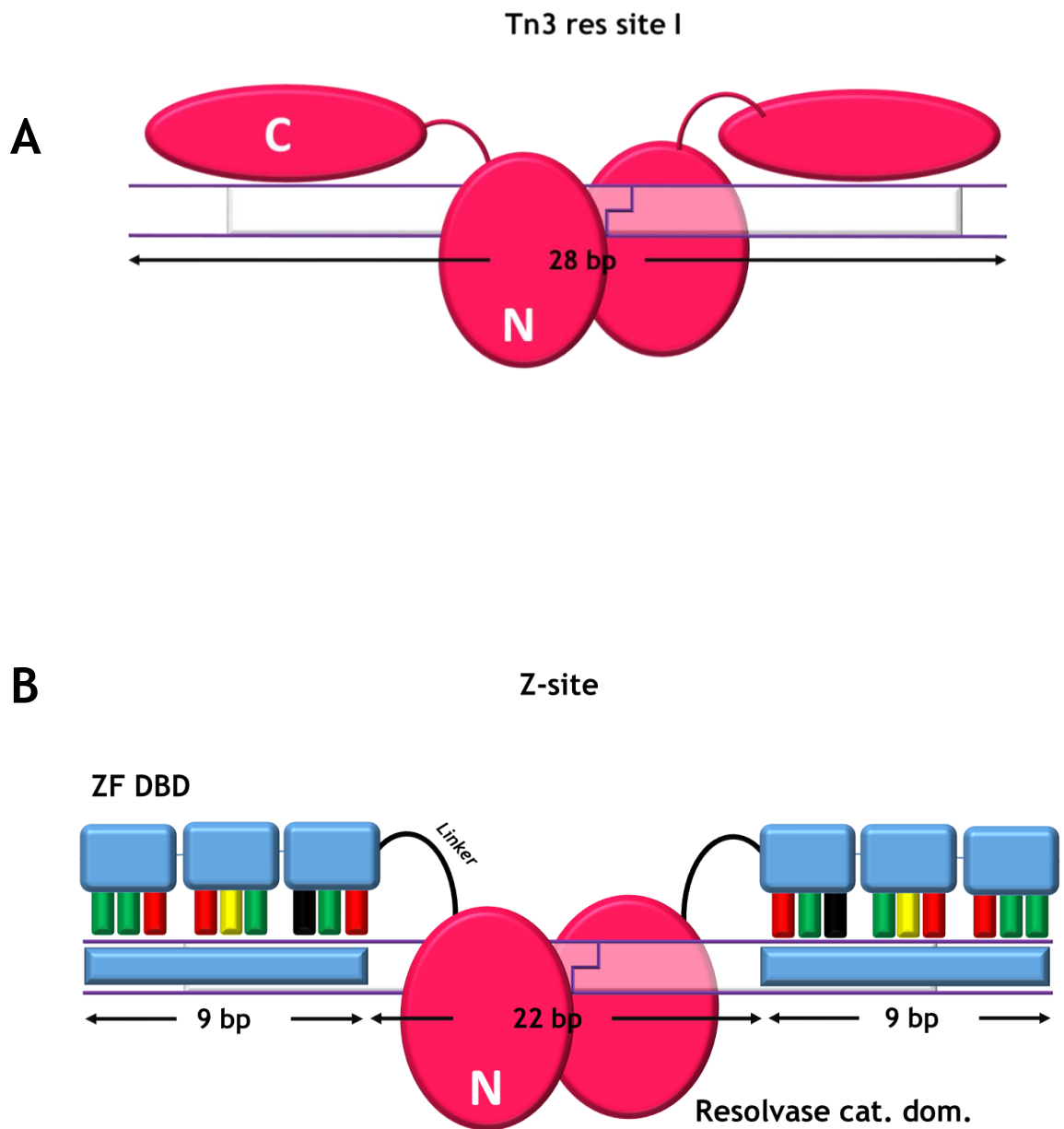


Figure 1.18: Zinc Finger Recombinases (ZFR) **A.** Cartoon representation of a Tn3 resolvase dimer on site I. **B.** Cartoon representation of a ZFR dimer on its Z-site. The C-terminal DNA-binding domain of Tn3 resolvase is replaced with a Zinc finger protein to redirect its activity to a modified sequence, the Z-site. The ZFR shows a highly modular system with a left recombinase, a right recombinase, a left Zinc Finger protein (ZFP) domain and a right ZFP. The ZFP and catalytic domains are joined by a short linker. The architectural similarity of Tn3 *res* site I and ZFR Z-site is depicted as well as the structural organization of dimeric Tn3 ZFRs to mimic Tn3 resolvase activity and function.

With the current status of ZFR technology, one cannot simply choose a target sequence within the genome for targeting; the pool defines the potential target sites (Section 1.3.1.1), and this does not provide the full programmability required for precision genome editing. The OPEN and CoDA pools do not contain fingers for all possible triplet nucleotide (Section 4.1). Although zinc finger recombinases have provided a solid proof of concept for the application of CRs for genome editing, the unique challenges of context dependency and limited zinc finger pools for novel zinc finger protein design beg for the design and utilization of a fully programmable DNA-binding module for targeting the recombinase catalytic domain. Transcription activator-like effector (TALE) domains seem to be just the tool for this.

1.9.2 TALE Recombinases

In 2012, Mercer and colleagues reported on the design of an active TALE-recombinase (TALER). They generated a TALER fusion protein between AvrXa7, a close relative of PthXo1 with 94% sequence similarity (Section 1.3.1.2); and a mutant version of the Gin invertase, a serine recombinase which they have used previously to generate active ZFRs for bacterial and mammalian recombination (Gaj *et al.*, 2011). Unlike the architecture of TALENs (Section 1.3.1.2), the catalytic domain of the recombinase has to be fused to the N-terminal end of the TALE domain to allow for the formation of a synaptic complex. To accommodate this change in design, they first defined the optimal TALER target site orientation showing that having the TALE binding sequence on the sense strand opposite the recombination crossover site yielded an active conformation (Fig. 1.19). They used a bacterial plasmid-based split gene assay where recombinase-mediated excision of a GFP insert resulted in the reconstitution of the β -lactamase gene, conferring ampicillin resistance. Surviving clones on ampicillin plates were expected to contain active TALERs and to have lost GFP fluorescence.

Using an incremental truncation assay based on exonuclease digestion and blunt ended cloning, they generated a library of N-terminal TALE truncated mutants (ranging from residues 1 to 298 of AvrXa7) and proposed that N-terminal deletions

between positions 120 and 129 were most stable for the generation of TALERs. Spacer length analysis indicated that a 32-bp core crossover 'gix' site was most favoured. The highest level of Gin-TALER mediated recombination reported in *E. coli* here was 7%. They also demonstrated TALER-mediated recombination activity in human embryonic kidney (HEK) 293T cells by measuring fold-reduction of luciferase expression. Results from a mammalian cell-based assay showed considerable differences from those obtained in *E. coli*. The mammalian luciferase-based assay showed very low recombination activity and reasons for this were not provided.

While this work has yielded significant information on the design of TALERs and their target sites, the level of recombination shown in both the bacterial and mammalian systems does not present convincing evidence of the potential of TALERs for significant clinical genome editing work. The efficiency of these systems is just too low relative to what current nuclease-based systems provide.

Concomitantly, research work on TALERs was also being carried out in our laboratory by Stephanie Holt (Holt, 2014) (Fig. 1.19). Here a PthXo1-based dTALEN, TALEN1297, one of a TALEN pair designed to target the *hey2* gene in Zebra fish (Sander *et al.*, 2011), formed the basis of the design. The N-terminal domain TALE1297 was modified to generate $\Delta 48$, $\Delta 84$, $\Delta 119$ and $\Delta 149$ truncations. The C-terminal Fok1 nuclease domain was also removed. N-terminal fusion proteins with the Tn3 resolvase deregulated mutant, NM resolvase, were generated and tested for excision activity against sites with 20, 22, 24 and 26-bp spacer lengths (lengths of the central sequence between TALE-binding sites). These were assayed in *E. coli* using a colorimetric MacConkey agar-based assay (Akopian *et al.*, 2003). No recombination activity was observed on these constructs. The four proteins were purified and assayed on the same sites. Again, no recombination activity was observed; however, the $\Delta 149$ TALER showed cleavage activity on the 22-bp and 24-bp spaced sites. This implied that the protein was able to cut the DNA but not to religate the cleaved ends.

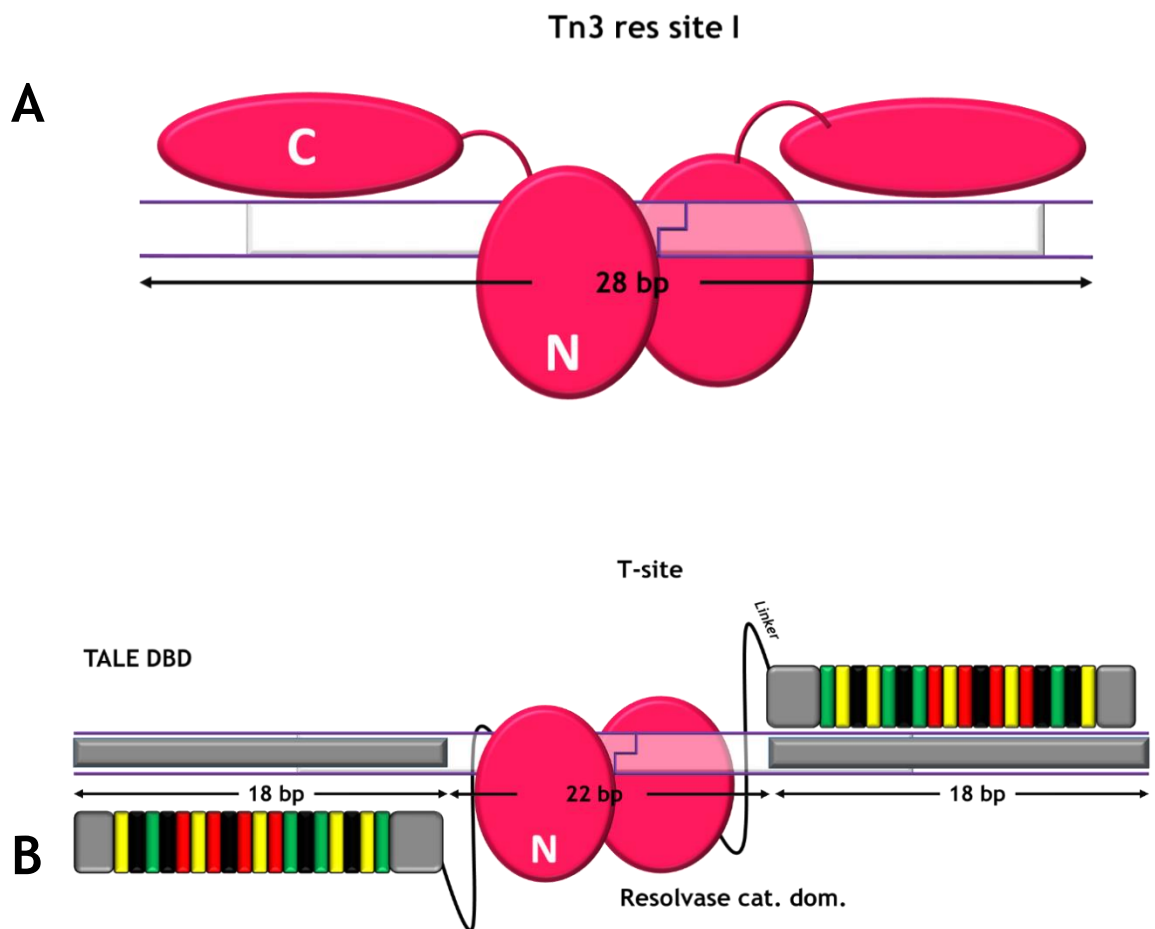


Figure 1.19: TALE Recombinases (TALER) A. Cartoon representation of a Tn3 resolvase dimer on site I B. Cartoon representation of a TALER dimer on its putative T-site. The C-terminal DNA binding domain of Tn3 resolvase is replaced with a TALE protein to redirect its activity to a modified sequence, the T-site. The sequence organization of T-sites is different from that of Tn3 resolvase on site I or ZFRs on Z-sites. This design change occurs because the resolvase catalytic domain is fused to the N-terminal domain of the TALE protein and the TALE protein binds DNA in the 5' to 3' direction from N- to C- terminus. This T-site design is the putative optimal design for Tn3 TALER as defined by Holt, (2014).

Electrophoretic mobility shift assays also showed what appeared to be a tetrameric complex of the $\Delta 149$ TALER bound to DNA. The conclusion that this work came to is that the formation of a stable synaptic complex is being impeded by the wrong positioning of the catalytic domains relative to the TALER target sites (T-sites) (Holt, 2014). The TALERS designed here serve as the starting point for the design of functional TALERS for HIV proviral DNA excision described in this work.

To date, minimal activities have been demonstrated by the TALERS, bringing up the need for optimization and questions about the effect of the TAL-effector domain on the recombinase catalytic ability.

1.9.3 RNA-guided Recombinases

Recently, a Gin recombinase-CRISPR/Cas9 complex has been reported that carries out recombination with up to 32% efficiency in mammalian systems (Yang *et al.*, 2017; Chaikind *et al.*, 2016).

This CRISPR-recombinase, recCas9 was generated by fusing catalytically inactive dCas9 with a hyperactive Gin recombinase catalytic domain variant (Gin β) (Gaj *et al.*, 2013) that demonstrates broadened recombination specificity for several 20-bp 'gix' sequences (Chaikind *et al.*, 2016). recCas9 was targeted to multiple endogenous loci in HEK 293T cells, catalysing seamless cleavage and religation with efficiency ranging between 12% to 32%. Off-target mutations were not assayed for in this work. *In silico* site search with consideration of the PAM requirement and gix sequences revealed that recCas9 could be targeted to 450 potential sites in the human genome. This broadened specificity of the Gin β catalytic domain piggybacked on dCas9 in the recCas9 construct could lead to devastating results in clinical applications.

Next-generation RNA-guided recombinases could utilize reprogrammed recombinase catalytic domains that do not have inherent promiscuous activity. It is also important to note here that recombinases are highly flexible proteins and their fusion to a large protein like Cas9 as done in this work could limit the essential conformational changes required for activity; smaller Cas9 orthologs like Cpf1 might provide more favourable results in CRISPR-recombinase design.

In spite of the significant amount of information and computational tools for the design of CRISPR systems and the high efficiency the tools offers off-target activities stand against their full adoption for clinical applications (Peng *et al.*, 2016). For example, Liang *et al.* (2015) show alarming off-target effects in the use of CRISPR/Cas9 for editing human tripronuclear (3pn) zygotes. The requirement of PAM sequences might also limit the flexibility of target site selection although the influx of newer Cas gene variants and their PAMs might surmount this challenge soon (Leenay *et al.*, 2016; Karvelis *et al.*, 2015). Also, the current reliance on HDR and NHEJ for repair of target sites presents a huge challenge that the CRISPR/dCas9 recombinase system can solve. However, the mixed complexities of PAM requirements and recombinase catalytic domain sequence specificity might dissuade would-be adopters of this tool.

1.10 Research Aim

This research work aims to engineer chimaeric recombinases capable of excising HIV-1 proviral DNA from the infected genome, thereby, providing a potential cure (Fig. 1.20). Genome editing has been shown to be a viable alternative to currently approved ART. It has also been discussed that the currently available nuclease-based tools for this do not provide the level of safety, precision and efficiency essential for clinical applications. Different variants of chimaeric recombinases have also been reported, all with varying levels of efficiency and activity and most just at the proof-of-concept phase.

Instead of scouting for favourable endogenous loci for targeting, a highly conserved region of the HIV-1 LTR, the TATA box, which is consistent across several clades will be targeted (Section 1.2). This sequence is already present in direct repeat within the cDNA; as such incorporated provirus can be excised leaving behind a non-functional small scar sequence.

For HIV-1 proviral excision, chimaeric recombinases will not only mediate cleavage as in the strategies involving nucleases, but will also re-join the DNA, optimising precision and reducing the risks associated with NHEJ and incomplete HDR. The requirement of two sites for recombination catalysis will also largely prevent off-target activities presenting significant precision, robustness and high fidelity for genome editing applications.

A modular engineering approach is taken in this work. First the catalytic domain of Tn3 NM resolvase was re-engineered to target the core HIV resolvase target sequence using a combination of rational and random mutagenesis- this is covered exhaustively in Chapter 3. Then, the optimal functional architecture and properties of the Tn3 TALER are defined in Chapter 4. The two domains are then combined in Chapter 5 to generate full HIV TATA-targeting TALERs that catalyse excision of a mock-HIV template. An overview of the materials and methods utilized in this project is provided in Chapter 2.

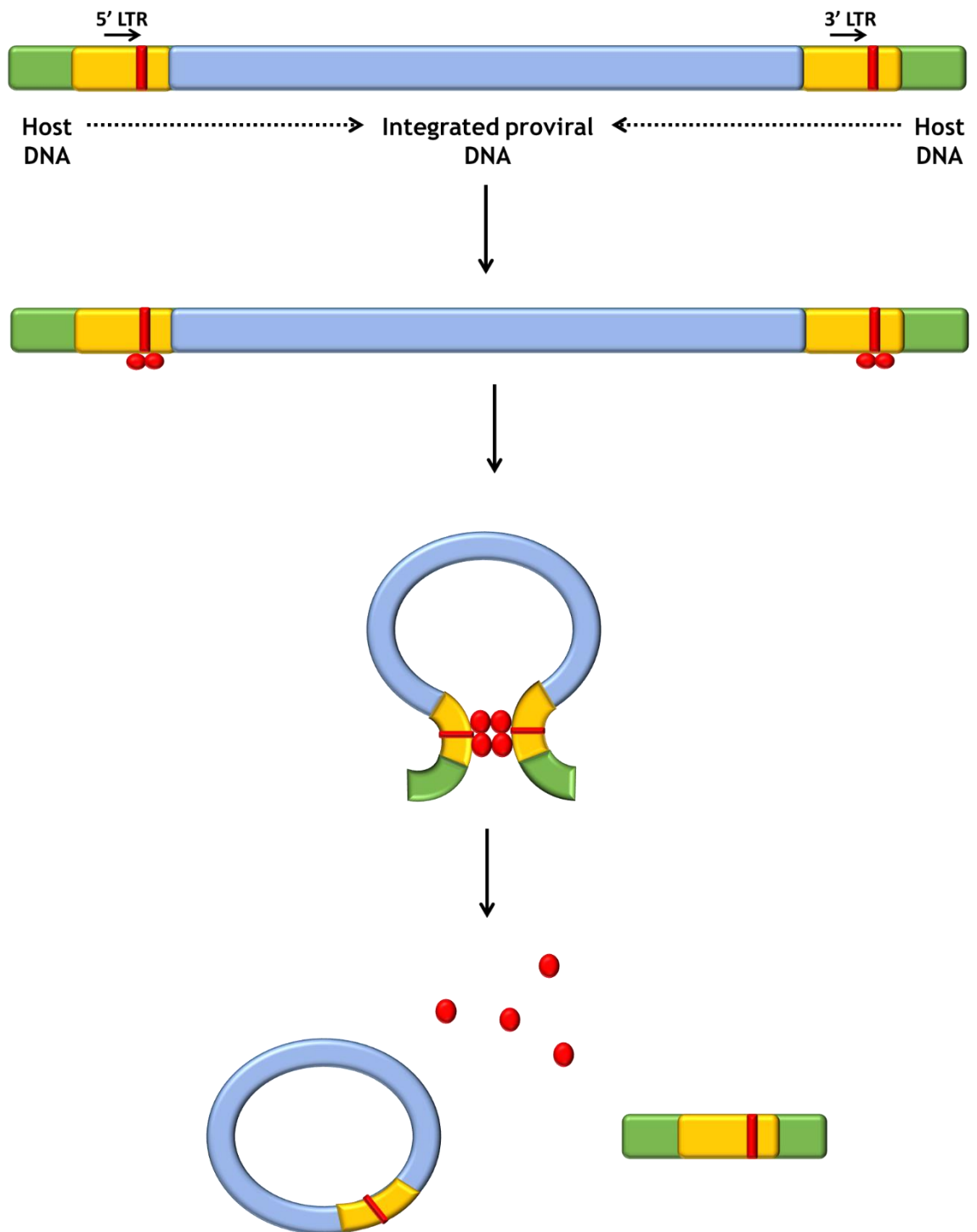


Figure 1.20: Targeted HIV-1 proviral excision from infected genome. Here, the programmable capacity of recombinases is harnessed in the modular engineering of CRs capable of recognizing a short HIV target sequence to catalyse excision. Red ovals represent the CR, the blue segment is the HIV provirus flanked by the yellow regions which are the LTRs. The red band is the CR target sequence and the outlying green regions denote the host genome. The goal is targeted genomic excision of the integrated HIV-1 provirus from the infected host genome presenting a potential cure for the HIV-1 infection.

Chapter 2: Materials and Methods

2.1 Bacterial Strains

The bacterial strains used in this work are derivatives of *Escherichia coli* K-12 and B strains. Strain names, genotypes, and sources are provided in Table 2.1. The DH5 and DS941 strains were used for routine plasmid cloning and the BL21 DE3 pLysS strain was used for protein expression.

Table 2.1: Bacterial strains

Strain	Genotype	Source
DH5	<i>E. coli</i> K12 strain, F-/endA1, hsdR17 (rk-, mk+), supE44, thi-1, D(lacZYA -argF)U169, deoR, recA1, phoA, gyrA96, relA1, λ^-	Invitrogen
DS941	<i>E. coli</i> K12 strain AB1157, but recF143, supE44, lacZDM15, lacIq	Summers and Sherratt, 1988
BL21 DE3 pLysS	<i>E. coli</i> B strain, F ⁻ ompT gal dcm lon hsdS _B (r _B ⁻ m _B ⁻) λ (DE3 [lacI lacUV5-T7p07 ind1 sam7 nin5]) [malB ⁺] _{K-12} (λ^S) pLysS[T7p20 ori _{p15A}](Cm ^R)	Studier <i>et al.</i> , 1990

2.2 Chemicals

The chemicals used in this work were sourced from the companies listed in Table 2.2. All solutions were prepared with distilled water unless otherwise stated.

Table 2.2: Chemicals

Chemical	Source
General chemicals, biochemicals, organic solvents	Sigma-Aldrich, BDH, Thermo Fisher Scientific
Agarose, acrylamide	Invitrogen
Restriction enzymes buffers	New England Biolabs (NEB)
Ligase buffer NEB	New England Biolabs (NEB)
Growth Media	Difco

2.3 Bacterial growth media

Routine liquid cultures of *E. coli* were grown in lysogeny broth (L-broth) and solid cultures on L-agar (L-Broth with 15 g/l agar). These were sterilised at 120 °C for 15 minutes. MacConkey agar (supplemented with galactose) was used for *in vivo* recombination assays. To make the MacConkey agar plates, 8 g of MacConkey agar base is completely dissolved by swirling in 180 mL of distilled water and heating in a microwave oven. The hot solution was cooled to ~60 °C, after which 10 ml of

20% galactose (to 1% w/v) is added, as well as appropriate volumes of kanamycin and ampicillin. The mixture was gently swirled, poured into 8 petri dishes and then allow to set for about 30 minutes before drying in a Unitemp drying cabinet (LTE Scientific). Media composition is provided in Table 2.3.

Table 2.3: Bacterial growth media

Growth media	Composition
L-broth	10 g bacto-tryptone, 5 g bacto-yeast extract and 5 g NaCl, made up to 1 Litre using distilled water and pH adjusted to 7.5 with NaOH
L-agar	L-broth with 15 g/l agar
MacConkey agar base	17 g bacto-peptone, 3 g bactoprotease peptone, 1.5 g bacto bile salts No.3, 5 g NaCl, 13.5 g bacto agar, 0.03 g neutral red, 0.001 g bacto crystal violet

2.4 Antibiotics

Stock concentrations of antibiotics were usually prepared for storage as described in Table 2.4. For selective conditions in liquid and solid growth media, antibiotics were added up to the working concentration.

Table 2.4: Antibiotics

Antibiotic	Stock concentration (100X)	Working concentration
Ampicillin (Amp)	10 mg/ml in H ₂ O	100 µg /ml
Kanamycin (Km)	5 mg/ml in H ₂ O	50 µg /ml
Chloramphenicol (Cm)	2.5 mg/ml in ethanol	25 µg /ml

2.5 Custom DNA synthesis

Simple oligonucleotides for plasmid construction (Section 2.7) and fluorescently-labelled oligonucleotides for electrophoretic mobility shift assays (EMSA) (Section 2.14.3) were ordered from Eurofins Genomics (Table 2.5). Double-stranded DNA fragments ('gBlocks') for plasmid modification were ordered from Integrated DNA Technologies (IDT). Plasmids for HIV-TALER construction were constructed and manufactured by Thermo Fisher Scientific (Section 5.2.2).

2.6 Plasmids

The plasmids that were used and designed in this work are listed in Table 2.6.

Table 2.5: List of custom oligonucleotides and synthetic DNA used in this study

Name	Size (bases)	Sequence (5' to 3')	Purpose
M13 uni (-43)	23	AGGGTTTTCCCAGTCACGACGTT	Sequencing primer
M13 rev (-49)	24	GAGCGGATAACAATTTACACACAGG	Sequencing primer
T7	20	TAATACGACTCACTATAGGG	Sequencing primer
site-A141(F)	27	TCATGATGATATATTTTTATCTTGTGC	Sequencing primer
site-B141(F)	18	AGGTGGCGTACGCATGAC	Sequencing primer
Site-A (hivL-sin) top	47	CTAGTGCGTGGGCGAGCGCTGCATATATGCAGCAGCCGCCACGCGC	HIV-ZLL recombination site .
Site-A(hivL-sin)bot	47	GGCCGCGCGTGGGCGGCTGCTGCATATATGCAGCGCTCGCCCACGCA	HIV-ZLL recombination site.
Site-B(hivL-sin)top	50	AATTCGCGTGGGCGAGCGCTGCATATATGCAGCAGCCGCCACGCGAGCT	HIV-ZLL recombination site .
Site-B(hivL-sin)bot	42	CGCGTGGGCGGCTGCTGCATATATGCAGCGCTCGCCCACGCG	HIV-ZLL recombination site.
Site-A(hivR-sin)top	47	CTAGTGCGTGGGCGAGCGCTGCTTATAAGCAGCAGCCGCCACGCGC	HIV-ZRR recombination site.
Site-A(hivR-sin)bot	47	GGCCGCGCGTGGGCGGCTGCTGCTTATAAGCAGCGCTCGCCCACGCA	HIV-ZRR recombination site.
Site-B(hivR-sin)top	50	AATTCGCGTGGGCGAGCGCTGCTTATAAGCAGCAGCCGCCACGCGAGCT	HIV-ZRR recombination site.
Site-B(hivR-sin)bot	42	CGCGTGGGCGGCTGCTGCTTATAAGCAGCGCTCGCCCACGCG	HIV-ZRR recombination site.

Names are highlighted in different colours according to research purpose.

Grey: Sequencing primers

Green: Recombination site oligonucleotides

Pink: TALER linker oligonucleotides

Blue: ZFR oligonucleotides for introducing mutations

Purple: Double-stranded DNA (gBlock)

Table 2.5- continued

Name	Size (bases)	Sequence (5' to 3')	Purpose
Site-A(hivLR)top	47	CTAGTGCGTGGGCGAGCGCTGCATATAAGCAGCAGCCGCCACGCGC	HIV-ZLR recombination site.
Site-A(hivLR)bot	47	GGCCGCGCGTGGGCGGCTGCTGCTTATATGCAGCGCTCGCCACGCA	HIV-ZLR recombination site.
Site-B(hivLR)top	50	AATTCGCGTGGGCGAGCGCTGCATATAAGCAGCAGCCGCCACGCGAGCT	HIV-ZLR recombination site.
Site-B(hivLR)bot	42	CGCGTGGGCGGCTGCTGCTTATATGCAGCGCTCGCCACGCG	HIV-ZLR recombination site.
Tn3LP_T	50	AATTCGCGTGGGCGAGCAAATATTATAATATTTAGCCGCCACGCGAGCT	Tn3LP recombination Z-site
Tn3LP_B	42	CGCGTGGGCGGCTAAATATTATAATATTTGCTCGCCACGCG	Tn3LP recombination Z-site
Tn3LP-3a_T	50	AATTCGCGTGGGCGAGCAAATAATATATTTAGCCGCCACGCGAGCT	Tn3LPA recombination Z-site
Tn3LP-3a_B	42	CGCGTGGGCGGCTAAATAATATATTTAGCCGCCACGCG	Tn3LPA recombination Z-site
Tn3LP-3g_T	50	AATTCGCGTGGGCGAGCAAATAGTATACTATTTAGCCGCCACGCGAGCT	Tn3LPG recombination Z-site
Tn3LP-3g_B	42	CGCGTGGGCGGCTAAATAGTATACTATTTGCTCGCCACGCG	Tn3LPG recombination Z-site
Tn3LP-3c_T	50	AATTCGCGTGGGCGAGCAAATACTATAGTATTTAGCCGCCACGCGAGCT	Tn3LPC recombination Z-site
Tn3LP-3c_B	42	CGCGTGGGCGGCTAAATACTATAGTATTTGCTCGCCACGCG	Tn3LPC recombination Z-site
Tn3RP_T	50	AATTCGCGTGGGCGAGCATAAATTTATAAATTATAGCCGCCACGCGAGCT	Tn3RP recombination Z-site
Tn3RP_B	42	CGCGTGGGCGGCTATAAATTTATAAATTATGCTCGCCACGCG	Tn3RP recombination Z-site
Tn3RP-3a_T	50	AATTCGCGTGGGCGAGCATAAATATATATATTATAGCCGCCACGCGAGCT	Tn3RPA recombination Z-site
Tn3RP-3a_B	42	CGCGTGGGCGGCTATAAATATATATATTATGCTCGCCACGCG	Tn3RPA recombination Z-site
Tn3RP-3g_T	50	AATTCGCGTGGGCGAGCATAATGTATACATTATAGCCGCCACGCGAGCT	Tn3RPG recombination Z-site
Tn3RP-3g_B	42	CGCGTGGGCGGCTATAATGTATACATTATGCTCGCCACGCG	Tn3RPG recombination Z-site
Tn3RP-3c_T	50	AATTCGCGTGGGCGAGCATAATCTATAGATTATAGCCGCCACGCGAGCT	Tn3RPC recombination Z-site
Tn3RP-3c_B	42	CGCGTGGGCGGCTATAATCTATAGATTATGCTCGCCACGCG	Tn3RPC recombination Z-site
HfTsAB_T	68	AATTCGGGAGTGGCCAACCCTCAGATGCTGCATATAAGCAGCTGCTTTTCGCCTGTACTGGGTGAGCT	HIV58T recombination T-site
HfTsAB_B	60	CACCCAGTACAGGCGAAAAGCAGCTGCTTATATGCAGCATCTGAGGGTTGGCCACTCCCG	HIV58T recombination T-site

Table 2.5- continued

Name	Size (bases)	Sequence (5' to 3')	Purpose
NMTsAA_T	68	AATTCGGGAGTGGCCAACCCTCAGATATAATTTATAATATTTTGCTGAGGGTTGGCCACTCCC GAGCT	HTA-Tn3-HTA recombination T-site
NMTsAA_B	60	CGGGAGTGGCCAACCCTCAGCAAATATTATAAATTATATCTGAGGGTTGGCCACTCCCG	HTA-Tn3-HTA recombination T-site
NMTsBB_T	68	AATTCACCCAGTACAGGCGAAAAGATATAATTTATAATATTTTGCTTTTCGCCTGTACTGGGT GAGCT	HTB-Tn3-HTB recombination T-site
NMTsBB_B	60	CACCCAGTACAGGCGAAAAGCAAATATTATAAATTATATCTTTTCGCCTGTACTGGGTG	HTB-Tn3-HTB recombination T-site
NMTsAB_T	68	AATTCGGGAGTGGCCAACCCTCAGATATAATTTATAATATTTTGCTTTTCGCCTGTACTGGGT GAGCT	HTA-Tn3-HTB recombination T-site
NMTsAB_B	60	CACCCAGTACAGGCGAAAAGCAAATATTATAAATTATATCTGAGGGTTGGCCACTCCCG	HTA-Tn3-HTB recombination T-site
HtHey2_T	68	AATTCAGATGTGGAAACGGAAGAGATGCTGCATATAAGCAGCTGCTCTTCCGTTTCCACATCT GAGCT	HIV-TLR Recombination T-site
HtHey2_B	60	CAGATGTGGAAACGGAAGAGCAGCTGCTTATATGCAGCATCTCTTCCGTTTCCACATCTG	HIV-TLR Recombination T-site
LinT12_T	43	GGCCGCAGGCGTACCGTGGACAGGGGCTCTGGCGGTTCCGGCA	6-aa GSGGSG linker for TALER6 optimization
LinT12_B	43	CTAGTGCCGGAACCGCCAGAGCCCCTGTCCACGGTACGCCTGC	6-aa GSGGSG linker for TALER6 optimization
LinT13_T	46	GGCCGCAGGCGTACCGTGGACAGGGGCGGTGGTTCTGGCGGCGGTA	7-aa GGGSGGG linker for TALER6 optimization
LinT13_B	46	CTAGTACCGCCGCCAGAACCACCGCCCCTGTCCACGGTACGCCTGC	7-aa GGGSGGG linker for TALER6 optimization
LinT14_T	52	GGCCGCAGGCGTACCGTGGACAGGGGCTCTGGCGGTTCCGGCGGCTCTGGTA	9-aa GSGGSGGSG linker for TALER6 optimization
LinT14_B	52	CTAGTACCAGAGCCGCCGGAACCGCCAGAGCCCCTGTCCACGGTACGCCTGC	9-aa GSGGSGGSG linker for TALER6 optimization

Table 2.5- continued

Name	Size (bases)	Sequence (5' to 3')	Purpose
LinT15_T	61	GGCCGCAGGCGTACCGTGGACAGGGGCTCTGGCGGTTCCGGCGGCTCTGGTGGCAGTGGTA	12-aa GSGGSGGSGGSG linker for TALER6 optimization
LinT15_B	61	CTAGTACCACTGCCACCAGAGCCGCCGGAACCGCCAGAGCCCCTGTCCACGGTACGCCTGC	12-aa GSGGSGGSGGSG linker for TALER6 optimization
JRig-link_T	61	GGCCGCAGGCGTACCGTGGACAGGGCCGAAGCTGCGGCAAAGAAGCAGCGGCTAAAGCCA	12-aa AEAAAKEAAKA linker for TALER6 optimization
JRig-link_B	61	CTAGTGGCTTTAGCCGCTGCTTCTTTTGCCGAGCTTCGGCCCTGTCCACGGTACGCCTGC	12-aa AEAAAKEAAKA linker for TALER6 optimization
009_T	60	GATCCTAGAGCGCACGAATGAGGGCAGACAGGCAGCAAAGCTTAAGGGAATCAAATTTGG	To introduce mutations present in pJUM009
009_B:	60	TCGACCAAATTTGATTCCCTTAAGCTTTGCTGCCTGTCTGCCCTCATTCGTGCGCTCTAG	To introduce mutations present in pJUM009
010_T:	41	GATCCTAGAGCGCACGAATGAGGGCAGACAGGAAGCAAAGC	To introduce mutations present in pJUM010
010_B:	41	TTAAGCTTTGCTTCCCTGTCTGCCCTCATTCGTGCGCTCTAG	To introduce mutations present in pJUM010
011_T:	72	TGAGCGCCGAGGATCCTAGAGCGCACGAATGAGGGCAGACAGGCAGCAAAGCTTAAGGGAGTCAAATTTGG	To introduce mutations present in pJUM011
011_B:	73	TCGACCAAATTTGACTCCCTTAAGCTTTGCTGCCTGTCTGCCCTCATTCGTGCGCTCTAGGATCCTCCGGCGC	To introduce mutations present in pJUM011
ZFR013_T:	72	TGAGCGCTTGAGGATCCTACAGCGCACGAATGAGGGCAGACAGGCAGCAAAGCTTAAGGGAGTCAAATTTGG	To introduce mutations present in pJUM006
ZFR013_B:	73	TCGACCAAATTTGACTCCCTTAAGCTTTGCTGCCTGTCTGCCCTCATTCGTGCGCTGTAGGATCCTCAAGCGC	To introduce mutations present in pJUM006
ZFR014_T:	72	TGAGCGCTTGAGGATCCTAGAGCGCACGAATGAGGGCAAGCAGGCAGCAAAGCTTAAGGGAGTCAAATTTGG	To introduce mutations present in pJUM007
ZFR014_B:	73	TCGACCAAATTTGACTCCCTTAAGCTTTGCTGCCTGTCTGCCCTCATTCGTGCGCTCTAGGATCCTCAAGCGC	To introduce mutations present in pJUM007

Table 2.5- continued

Name	Size (bases)	Sequence (5' to 3')	Purpose
ZFR017_T:	72	TGAGCGCTTGAGGATCCTAGAGCGCAAGAATGAGGGCTTACAGGCAGCAAAGCTTAAGGGAATCAAAGGTGG	To introduce mutations present in pJUM008
ZFR017_B:	73	TCGACCACCTTTGATTCCCTTAAGCTTTGCTGCCTGTAAGCCCTCATTCTTGCGCTCTAGGATCCTCAAGCGC	To introduce mutations present in pJUM008
051T1	30	CGCGTCTCAACCAGCCGGCAGTCCCTCGAT	To introduce mutations present in pJUM051
051T2	59	ATTCAGATCAGGGCGCTCAAAGATGCAGGGGTAGAAGCTAACCGCATCTTTACCGACAA	To introduce mutations present in pJUM051
051B1	50	ATCTTTGAGCGCCCTGATCTGAATATCGAGGGACTGCCGGCTGGTTGAGA	To introduce mutations present in pJUM051
051B2	39	AGCTTTGTGCGTAAAGATGCGGTTAGCTTCTACCCCTGC	To introduce mutations present in pJUM051
061_T	41	GATCCTAGAGCGCCTGAATGAGGGCAGACAGGCAGCAAAGC	To introduce T126L mutation into pJUM004 for future work
061_B	41	TTAAGCTTTGCTGCCTGTCTGCCCTCATTCAGGCGCTCTAG	To introduce T126L mutation into pJUM004 for future work
T126LIB_T	41	GATCCTAGAGCGCNSAATGAGGGCAGACAGGAAGCAAAGC	T126 single mutant library for future work on position -3
T126LIB_B	41	TTAAGCTTTGCTTCTGTCTGCCCTCATTSNNGCGCTCTAG	T126 single mutant library for future work on position -3
R130LIB_T	41	GATCCTAGAGCGCACGAATGAGGGCNSCAGGAAGCAAAGC	R130 single mutant library for future work on position -3
R130LIB_B	41	TTAAGCTTTGCTTCTGTSNNGCCCTCATTCGTGCGCTCTAG	R130 single mutant library for future work on position -3
Q131LIB_T	41	GATCCTAGAGCGCACGAATGAGGGCAGANNSGAAGCAAAGC	Q131 single mutant library for future work on position -3
Q131LIB_B	41	TTAAGCTTTGCTTCSNNTCTGCCCTCATTCGTGCGCTCTAG	Q131 single mutant library for future work on position -3

Table 2.5- continued

Name	Size (bases)	Sequence (5' to 3')	Purpose
K134LIB_T	41	GATCCTAGAGCGCACGAATGAGGGCAGACAGGAAGCANNSC	K134 single mutant library for future work on position -3
K134LIB_B	41	TTAAGSNNTGCTTCTGTCTGCCCTCATTCGTGCGCTCTAG	K134 single mutant library for future work on position -3
F140LIB_T	19	TTAAGGGAATCAAANNSGG	F140 single mutant library for future work on position -3
F140LIB_B	19	TCGACCSNNTTTGATTCCC	F140 single mutant library for future work on position -3
Hiscas_top	38	CACATCACCATCACCATCACTAATAAGCTAGCGGGTAC	To introduce 6x His-tag into ZFR over-expression plasmids
Hiscas_bot	38	CCGCTAGCTTATTAGTGATGGTGATGGTGATGTGAGCT	To introduce 6x His-tag into ZFR over-expression plasmids
Name	Size (bp)	Sequence (5' to 3')	Purpose
CtermTalGcas	300	TTTATCGCCTCGCAGATGGACAATTGCTAGCCACGACGGCGGACGGCCCGCCCTGGAGAGCAT TGTGGCCCAGCTGTCTAGACCTGATCCTGCCCTGGCCGCGTTAACGAATGACCATCTGGTGGC GTTGGCATGTCTTGGTGGACGACCCGCGCTCGATGCAGTCAAAAAGGGTCTGCCTCATGCTCC CGCATTGATCAAAAGAACCAACCGGCGAATTCCCGAGAGAACTTCCCACCGAGTGGCGCATCA CCATCACCATCACTGATGATCAGGTACCATCAAACAAGGACGCTAATC	To correct C-terminal truncation at the end of TALER100 and TALER101 and introduce 6x His-tag for protein purification.

Table 2.6: Plasmids used in this research work and their precursors.

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pMTL23	Ap ^r	2505	High copy-number cloning vector	Chambers <i>et al.</i> , 1988
pMS140	Ap ^r	5469	pAT5Δ derivative encoding the Tn3 resolvase ORF, with low level expression suitable for <i>in vivo</i> recombination assays	M.Stark
pMS183Δ	Km ^r	4863	<i>In vivo</i> recombination substrate precursor plasmid, allows cloning of recombination sites between BglII/BsrGI and NcoI/XbaI flanking <i>galk</i> indicator gene	M.Prorocic
pSA1101	Km ^r	6692	Wild type Tn3 resolvase expression plasmid with inducible T7 promoter	Arnold <i>et al.</i> , 1999
pFM141	Km ^r	4882	<i>In vivo</i> recombination substrate precursor plasmid, allows one-step cloning of annealed oligonucleotides recombination between SpeI/NotI and EcoRI/SacI flanking <i>galk</i> indicator gene	F. Olorunniji
pFM160	Ap ^r , Km ^r	4197	<i>In vitro</i> recombination substrate precursor plasmid, allows one-step cloning of annealed oligonucleotides recombination between SpeI/NotI and EcoRI/SacI flanking KmR marker	F. Olorunniji
pUC71K	Ap ^r , Km ^r	3914	Cloning vector carrying a Kanamycin cassette flanked by BamHI for generation of <i>in vitro</i> recombination substrate plasmid	Vieira and Messing, 1982
pFO2	Ap ^r	5469	pMS140 derivative encoding Tn3 resolvase mutant NM (R2A, E56K, G101S, D102Y, M103I, Q105L) ORF	F. Olorunniji
pFO32	Ap ^r	5469	pMS140 derivative encoding Tn3 resolvase mutant NM, S10A ORF	F. Olorunniji
pMP59	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR (R2A, E56K, G101S, D102Y, M103I, Q105L) ORF	M. Prorocic
pMP213	Ap ^r	5654	Low-level expression plasmid encoding Sin ZFR ORF	M. Prorocic

Unless otherwise indicated (in Source / Reference), all plasmids were constructed in this study. ≥ 119 plasmids were constructed in this study.

Plasmid names are highlighted in different colours according to type and purpose.

Yellow: pMTL23-based single recombination site plasmids (Section 2.10.2)

Blue: pMS140-based low level expression plasmids (Section 2.10.1)

Orange: pSA1101-based over-expression plasmids (Section 2.10.1)

Green: Two-site *in vivo* recombination substrate plasmids (Section 2.10.2)

Purple: Two-site *in vitro* recombination substrate plasmids (Section 2.10.2)

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pCP591 (cMutA)	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L R120Q) ORF	C. Proudfoot
pCP599 (cMutB)	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (R120Q L135R) ORF	C. Proudfoot
pCP620 (cMutC)	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L E132A) ORF	C. Proudfoot
pCP655 (cMutD)	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (A115G R120Q E132A) ORF	C. Proudfoot
pCP657 (cMutE)	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L A115G R120Q E132A) ORF	C. Proudfoot
pCP734 (cMutF)	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (R120V R130I L135R I138V K139Y F140L) ORF	C. Proudfoot
pCP735 (cMutG)	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (R120F R130I E132A L135R K136R I138V K139Y F140N) ORF	C. Proudfoot
pCP739 (cMutH)	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (R120F R130I E132A K136R I138V K139Y F140L) ORF	C. Proudfoot
pAM0030	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (G70S) ORF	C. Proudfoot
pAM0031	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (V107L) ORF	C. Proudfoot
pAM0036	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (G70S V107L) ORF	C. Proudfoot
pAM0013	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L V107F)	C. Proudfoot
pAM0122	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E G70S V107L) ORF	C. Proudfoot
pMP53	Km ^r	4943	<i>In vivo</i> recombination substrate with two copies of Z22Z site in direct repeat flanking a <i>galk</i> gene, pMS183D backbone	M. Prorocic
pMP217	Km ^r	4978	<i>In vivo</i> recombination substrate with two copies of Z22Z (Sin) Z-site in direct repeat flanking a <i>galk</i> gene, pMS183D backbone	M. Prorocic
pMP243	Km ^r	4960	<i>In vivo</i> recombination substrate with two copies of Tn3 site I in direct repeat flanking a <i>galk</i> gene, pMS183D backbone	M. Prorocic

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pMP78	Ap ^r , Km ^r	3919	<i>In vitro</i> recombination substrate with two copies of Tn3 site I, flanking the KmR marker from pUC71K	M. Prorocic
pDWIVS6	Ap ^r , Km ^r		<i>In vitro</i> recombination substrate with two copies of Tn3 Z-site22, flanking the KmR marker from pUC71K	D. Wenlong/ M. Stark
pSE1021 (TS22)	Km ^r	4984	<i>In vivo</i> recombination substrate, with two copies of TALE1297-Tn3 T-site22 in direct repeat, flanking a <i>galk</i> gene	S. Holt
pSE1029 (TS24)	Km ^r	4988	<i>In vivo</i> recombination substrate, with two copies of TALE1297-Tn3 T-site24 in direct repeat, flanking a <i>galk</i> gene	S. Holt
pSE1030 (TS26)	Km ^r	4992	<i>In vivo</i> recombination substrate, with two copies of TALE1297-Tn3 T-site26 in direct repeat, flanking a <i>galk</i> gene	S. Holt
pSE1031 (TS28)	Km ^r	4996	<i>In vivo</i> recombination substrate, with two copies of TALE1297-Tn3 T-site28 in direct repeat, flanking a <i>galk</i> gene	S. Holt
pSE1042 (IVTS22)	Ap ^r , Km ^r	3959	<i>In vitro</i> recombination substrate with two copies of TALE1297-Tn3 T-site22 in direct repeat flanking the KmR marker from pUC71K	S. Holt
pSE1043 (IVTS24)	Ap ^r , Km ^r	3963	<i>In vitro</i> recombination substrate with two copies of TALE1297-Tn3 T-site24 in direct repeat flanking the KmR marker from pUC71K	S. Holt
pSE1044 (IVTS26)	Ap ^r , Km ^r	3967	<i>In vitro</i> recombination substrate with two copies of TALE1297-Tn3 T-site26 in direct repeat flanking the KmR marker from pUC71K	S. Holt
pSE1045 (IVTS28)	Ap ^r , Km ^r	3971	<i>In vitro</i> recombination substrate with two copies of TALE1297-Tn3 T-site28 in direct repeat flanking the KmR marker from pUC71K	S. Holt
pSE1020 (TIS 22)	Ap ^r	2563	pMTL23 with TALE1297-Tn3 T-site22 cloned between its EcoRI and SacI sites	S. Holt
pSE1024 (TIS 24)	Ap ^r	2565	pMTL23 with TALE1297-Tn3 T-site24 cloned between its EcoRI and SacI sites	S. Holt
pSE1025 (TIS 26)	Ap ^r	2567	pMTL23 with TALE1297-Tn3 T-site26 cloned between its EcoRI and SacI sites	S. Holt
pSE1026 (TIS 28)	Ap ^r	2569	pMTL23 with TALE1297-Tn3 T-site28 cloned between its EcoRI and SacI sites	S. Holt

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pSE2034 (LTE 1)	Ap ^r	7471	Low-level expression plasmid encoding TALER-221 with 6aa linker ORF (Tn3 NM resolvase R144-GSGGSG-TS-Δ221 TALE1297)	S. Holt
pSE2035 (LTE 2)	Ap ^r	7480	Low-level expression plasmid encoding TALER-221 with 9aa linker ORF (Tn3 NM resolvase R144-GSGGSGGSG-TS-Δ221 TALE1297)	S. Holt
pSE2036 (LTE 3)	Ap ^r	7489	Low-level expression plasmid encoding TALER-221 with 12aa linker ORF (Tn3 NM resolvase R144-GSGGSGGSGGSG-TS-Δ221 TALE1297)	S. Holt
pSE2037 (LTE 4)	Ap ^r	7474	Low-level expression plasmid encoding TALER-221 with 7aa linker ORF (Tn3 NM resolvase R144-GGGSGGG-TS-Δ221 TALE1297)	S. Holt
pSE2040 (LTE 5)	Ap ^r	7666	Low-level expression plasmid encoding TALER-154 ORF (Tn3 NM resolvase R144-TS-Δ154 TALE1297)	S. Holt
pSE2041 (LTE 6)	Ap ^r	7684	Low-level expression plasmid encoding TALER-154 with 6aa linker ORF (Tn3 NM resolvase R144-TS- GSGGSG-Δ154 TALE1297)	S. Holt
pSE2065 (LTE 7)	Ap ^r	8131	Low-level expression plasmid encoding TALER1 with 6aa linker ORF (Tn3 NM resolvase R144-TS- GSGGSG-1 TALE1297)	S. Holt
pSE2085 (LTE 8)	Ap ^r	7990	Low-level expression plasmid encoding TALER-48 with 6aa linker ORF (Tn3 NM resolvase R144-TS- GSGGSG-48 TALE1297)	S. Holt
pSE2095 (LTE 9)	Ap ^r	7882	Low-level expression plasmid encoding TALER-84 with 6aa linker ORF (Tn3 NM resolvase R144-TS- GSGGSG-84 TALE1297)	S. Holt
pSE2105 (LTE 10)	Ap ^r	7777	Low-level expression plasmid encoding TALER-119 with 6aa linker ORF (Tn3 NM resolvase R144-TS- GSGGSG-119 TALE1297)	S. Holt
pSE2115 (LTE 11)	Ap ^r	7687	Low-level expression plasmid encoding TALER-148 with 6aa linker ORF (Tn3 NM resolvase R144-TS- GSGGSG-84 TALE1297)	S. Holt
pJU001	Km ^r	4912	<i>In vivo</i> recombination substrate, with two copies of HIV-ZLL Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.2
pJU002	Km ^r	4912	<i>In vivo</i> recombination substrate, with two copies of HIV-ZRR Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.2
pJU003	Km ^r	4912	<i>In vivo</i> recombination substrate, with two copies of HIV-ZLR Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.2

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pJU004	Km ^r	4905	Precursor low-copy <i>in vivo</i> recombination substrate, with one copy of HIV-ZLL Z-site in front of a <i>galk</i> gene (in pFM141 backbone)	Section 3.3.2
pJU005	Km ^r	4905	Precursor low-copy <i>in vivo</i> recombination substrate, with one copy of HIV-ZRR Z-site in front of a <i>galk</i> gene (in pFM141 backbone)	Section 3.3.2
pJU006	Km ^r	4905	Precursor low-copy <i>in vivo</i> recombination substrate, with one copy of HIV-ZLR Z-site in front of a <i>galk</i> gene (in pFM141 backbone)	Section 3.3.2
pJU201	Ap ^r , Km ^r	4171	<i>In vitro</i> recombination substrate with two copies of HIV-ZLL Z-site in direct repeat flanking a KmR marker (cloned in one-step cloning using pFM160)	Section 5.2.1
pJU202	Ap ^r , Km ^r	4171	<i>In vitro</i> recombination substrate with two copies of HIV-ZRR Z-site in direct repeat flanking a KmR marker (cloned in one-step cloning using pFM160)	Section 5.2.1
pJU203	Ap ^r , Km ^r	4171	<i>In vitro</i> recombination substrate with two copies of HIV-ZLR Z-site in direct repeat flanking a KmR marker (cloned in one-step cloning using pFM160)	Section 5.2.1
pJTn3LP	Km ^r	4948	<i>In vivo</i> recombination substrate, with two copies of Tn3LP Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.7
pJTn3LPA	Km ^r	4948	<i>In vivo</i> recombination substrate, with two copies of Tn3LPA Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.7
pJTn3LPC	Km ^r	4948	<i>In vivo</i> recombination substrate, with two copies of Tn3LPC Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.7
pJTn3LPG	Km ^r	4948	<i>In vivo</i> recombination substrate, with two copies of Tn3LPG Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.7
pJTn3RP	Km ^r	4948	<i>In vivo</i> recombination substrate, with two copies of Tn3RP Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.7
pJTn3RPA	Km ^r	4948	<i>In vivo</i> recombination substrate, with two copies of Tn3RPA Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.7
pJTn3RPC	Km ^r	4948	<i>In vivo</i> recombination substrate, with two copies of Tn3RPC Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.7
pJTn3RPG	Km ^r	4948	<i>In vivo</i> recombination substrate, with two copies of Tn3RPG Z-site in direct repeat, flanking a <i>galk</i> gene	Section 3.3.7

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pJU501	Ap ^r	2563	pMTL23 with HTA-Tn3-HTA T-site ₂₂ cloned between its EcoRI and SacI sites	Section 5.2.2
pJU502	Ap ^r	2563	pMTL23 with HTA-Tn3-HTB T-site ₂₂ cloned between its EcoRI and SacI sites	Section 5.2.2
pJU503	Ap ^r	2563	pMTL23 with HTB-Tn3-HTB T-site ₂₂ cloned between its EcoRI and SacI sites	Section 5.2.2
pJU504	Ap ^r , Km ^r	4984	<i>In vitro</i> recombination substrate with two copies of HTA-Tn3-HTA T-site ₂₂ in direct repeat flanking the KmR marker from pUC71K (fragment inserted in orientation X with NruI restriction sites close together)	Section 5.2.2
pJU505	Ap ^r , Km ^r	4984	<i>In vitro</i> recombination substrate with two copies of HTA-Tn3-HTA T-site ₂₂ in direct repeat flanking the KmR marker from pUC71K (fragment inserted in orientation Y with NruI restriction sites far apart)	Section 2.10.2
pJU506	Ap ^r , Km ^r	4984	<i>In vitro</i> recombination substrate with two copies of HTA-Tn3-HTB T-site ₂₂ in direct repeat flanking the KmR marker from pUC71K (fragment inserted in orientation X with NruI restriction sites close together)	Section 5.2.2
pJU507	Ap ^r , Km ^r	4984	<i>In vitro</i> recombination substrate with two copies of HTA-Tn3-HTB T-site ₂₂ in direct repeat flanking the KmR marker from pUC71K (fragment inserted in orientation Y with NruI restriction sites far apart)	Section 2.10.2
pJU508	Ap ^r , Km ^r	4984	<i>In vitro</i> recombination substrate with two copies of HTB-Tn3-HTB T-site ₂₂ in direct repeat flanking the KmR marker from pUC71K (fragment inserted in orientation X with NruI restriction sites close together)	Section 5.2.2
pJU509	Ap ^r , Km ^r	4984	<i>In vitro</i> recombination substrate with two copies of HTB-Tn3-HTB T-site ₂₂ in direct repeat flanking the KmR marker from pUC71K (fragment inserted in orientation Y with NruI restriction sites far apart)	Section 2.10.2
pJU510	Ap ^r	2563	pMTL23 with HIV58T T-site ₂₂ cloned between its EcoRI and SacI sites	Section 5.2.3
pJU511	Ap ^r , Km ^r	4207	<i>In vitro</i> recombination substrate with two copies of HIV58T T-site ₂₂ in direct repeat (cloned using pFM160)	Section 5.2.3

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pJU512 IVTS 10	Ap ^r , Km ^r	4973	<i>In vitro</i> recombination two-site substrate plasmid with one copy of HIV-ZLL Z-site ²² and one copy of HIV58T T-site ²² in direct repeat flanking the KmR marker	Section 2.10.2
pJU513 IVTS 11	Ap ^r , Km ^r	4973	<i>In vitro</i> recombination two-site substrate plasmid with one copy of HIV-ZRR Z-site ²² and one copy of HIV58T T-site ²² in direct repeat flanking the KmR marker	Section 2.10.2
pJU514 IVTS 12	Ap ^r , Km ^r	4973	<i>In vitro</i> recombination two-site substrate plasmid with one copy of HIV-ZLR Z-site ²² and one copy of HIV58T T-site ²² in direct repeat flanking the KmR marker	Section 2.10.2
pJU515 HIV-TLR	Ap ^r , Km ^r	4984	<i>In vitro</i> recombination substrate with two copies of HIV-TLR T-site ²² (22-bp central HIV/TALE1297 T-target sequences) in direct repeat flanking the KmR marker from pUC71K	Section 5.2.1
pJU516	Ap ^r , Km ^r	4973	<i>In vitro</i> recombination two-site substrate plasmid with one copy of TALE1297-Tn3 T-site ²² and one copy of Tn3 Z-site ²² in direct repeat flanking the KmR marker	Section 4.2.3
pJU517	Ap ^r , Km ^r	4964	<i>In vitro</i> recombination two-site substrate plasmid with one copy of TALE1297-Tn3 T-site ²² and one copy of Tn3 <i>res</i> site I in direct repeat flanking the KmR marker	Section 4.2.3
pJU518	Ap ^r , Km ^r	4973	<i>In vitro</i> recombination two-site substrate plasmid with one copy of TALE1297-Tn3 T-site ²² and one copy of Sin Z-site ²² in direct repeat flanking the KmR marker	Section 4.2.3
pJU519	Ap ^r	4477	Derivative of pJU511 with <i>galk</i> gene replacing KmR gene. <i>In vitro</i> recombination substrate with two copies of HIV58T T-site ²² in direct repeat	Section 2.10.2
pJU550 (pJHFAB3)	Km ^r	4986	<i>In vivo</i> recombination substrate with two copies of HIV58T T-site ²² in direct repeat flanking the <i>galk</i> gene (HIV58T-L)	Section 5.2.3
pJUM001	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L R120D T126M R130I I138V K139N) ORF- selected from library screen	Section 3.3.5
pJUM003	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L R120I T126M L135R I138V F140M) ORF- selected from library screen	Section 3.3.5
pJUM004	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L R120L E132A I138V) ORF- selected from library screen	Section 3.3.5

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pJUM005	Ap ^r	11272	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L R120F L123I E132A F140N) ORF- selected from library screen as dimer plasmid	Section 3.3.5
pJUM005x	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L R120F L123I E132A F140N) ORF. Monomer variant of pJUM005	Section 3.3.5
pJUM006	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L R120L E124Q E132A I138V) ORF. Cloned into pJUM004 backbone using annealed oligonucleotides with BamHI and AflII silent restriction sites inserted for ease of future cloning.	Section 3.3.5
pJUM007	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L R120L R130K E132A I138V) ORF	Section 3.3.5
pJUM008	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L R120L T126K R130L E132A F140G) ORF	Section 3.3.5
pJUM009	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (R120L E132A) ORF	Section 3.3.5
pJUM010	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (R120L I138V) ORF	Section 3.3.5
pJUM011	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (E132A I138V) ORF	Section 3.3.5
pJUM012 (pCP556)	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E I77L V107F) ORF	C. Proudfoot/ Section 3.3.6
pJUM013 (pAM0013)	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L V107F) ORF	C. Proudfoot/ Section 3.3.6
pJUM014	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E I77L V107F R120L E132A I138V) ORF	Section 3.3.6
pJUM015	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E I77L R120L E132A I138V) ORF	Section 3.3.6
pJUM016	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (E46K V107F) ORF - (mutant p3mut5)	Section 3.3.7.3
pJUM017	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (V107F I138V) ORF - (mutant p3mut2)	Section 3.3.7.3

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pJUM018	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L I80V Q131R V107F) ORF - (mutant p3mut16)	Section 3.3.7.3
pJUM019	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L V107F) ORF - (mutant p3mut3)	Section 3.3.7.3
pJUM023	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (D25A I77L V107F) ORF - (mutant p3mut11)	Section 3.3.7.3
pJUM025	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant F34L I77L V107F) ORF - (mutant p3mut6)	Section 3.3.7.3
pJUM026	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (D25G V107L) ORF - (mutant p3mut9)	Section 3.3.7.3
pJUM027	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (V107L) ORF - (mutant p3mut7)	Section 3.3.7.3
pJUM028	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (V107F) ORF - (mutant p3mut1)	Section 3.3.7.3
pJUM029	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (Q13H V107F Q116L) ORF - (mutant p3mut14)	Section 3.3.7.3
pJUM031	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E V107F) ORF - (mutant p3mut4)	Section 3.3.7.3
pJUM032	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (Q13R V107F E128D) ORF - (mutant p3mut13)	Section 3.3.7.3
pJUM033	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (D25G I77L G87S V107F I138T) ORF - (mutant p3mut10)	Section 3.3.7.3
pJUM037	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (Q13R V107F) ORF - (mutant p3mut12)	Section 3.3.7.3
pJUM038	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (F83L V107L) ORF - (mutant p3mut8)	Section 3.3.7.3
pJUM039	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I3T Q13L V63G F83L V107F) ORF - (mutant p3mut15)	Section 3.3.7.3

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pJUM041	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E I77L V107L) ORF.	Section 3.3.8
pJUM042	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (I77L V107L R120L E132A I138V) ORF	Section 3.3.8
pJUM043	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E I77L V107L R120L E132A I138V) ORF	Section 3.3.8
pJUM044	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (Q13R I77L R120L E132A I138V) ORF	Section 3.3.8
pJUM045	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (Q13R I77L V107F R120L E132A I138V) ORF	Section 3.3.8
pJUM046	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E F83L V107L) ORF	Section 3.3.8
pJUM047	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E F83L V107L R120L E132A I138V) ORF	Section 3.3.8
pJUM048	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (Q13H I77L V107L Q116L) ORF	Section 3.3.8
pJUM049	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (Q13H I77L R120L E132A I138V) ORF	Section 3.3.8
pJUM050	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E I77L I80V V107F Q131R) ORF	Section 3.3.8
pJUM051	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (Q13R K29E I77L R120L E132A I138V) ORF	Section 3.3.9
pJUM052	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (Q13R K29E I77L V107F R120L E132A I138V) ORF	Section 3.3.9
pJUM056	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (K29E I77L V107F R120I T126M L135R I138V F140M) ORF	Section 3.3.9
pJUM057	Ap ^r	5636	Low-level expression plasmid encoding Tn3 NM ZFR mutant (Q13R I77L R120I T126M L135R I138V F140M) ORF	Section 3.3.9

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pJUM201	Km ^r	6859	Over-expression plasmid of ZR001: Tn3 NM ZFR mutant (I77L R120D T126M R130I I138V K139N)	Section 3.3.5
pJUM203	Km ^r	6859	Over-expression plasmid of ZR003: Tn3 NM ZFR mutant (I77L R120I T126M L135R I138V F140M)	Section 3.3.5
pJUM204	Km ^r	6859	Over-expression plasmid of ZR004: Tn3 NM ZFR mutant (I77L R120L E132A I138V)	Section 3.3.5
pJUM204H	Km ^r	6859	Over-expression plasmid of ZR004 with C-terminal 6x His-tag cassette in ORF	Section 3.3.5
pJUM205	Km ^r	6885	Over-expression plasmid of ZR005: Tn3 NM ZFR mutant (I77L R120F E132A F140N)	Section 3.3.6
pJUM212	Km ^r	6859	Over-expression plasmid of ZR012: Tn3 NM ZFR mutant (K29E I77L V107F)	Section 3.3.6
pJUM212H	Km ^r	6885	Over-expression plasmid of ZR012 with C-terminal 6x His-tag cassette in ORF	Section 3.3.6
pJUM213	Km ^r	6859	Over-expression plasmid of ZR113: Tn3 NM ZFR mutant (I77L V107F)	Section 3.3.6
pJUM213H	Km ^r	6885	Over-expression plasmid of ZR113 with C-terminal 6x His-tag cassette in ORF	Section 3.3.6
pJUM245H	Km ^r	6859	Over-expression plasmid of ZR113 with C-terminal 6x His-tag cassette in ORF. Tn3 NM ZFR mutant (Q13R I77L V107F R120L E132A I138V)	Section 5.2.1
pJUM410	Ap ^r	7702	Low level-expression plasmid encoding TALER6 ORF (R148-6aa linker- TS- D148)	Section 4.2.1
pJUM411	Ap ^r	7705	Low level-expression plasmid encoding TALER7 ORF (R148-7aa linker- TS- D148)	Section 4.2.1
pJUM412	Ap ^r	7711	Low level-expression plasmid encoding TALER9 ORF (R148-9aa linker- TS- D148)	Section 4.2.1
pJUM413	Ap ^r	7720	Low level-expression plasmid encoding TALER12 ORF (R148-12aa linker- TS- Δ148)	Section 4.2.1
pJUM414	Ap ^r	7684	Low level-expression plasmid encoding TALER0 ORF (R148- TS- Δ148)	Section 4.2.1
pJUM415	Ap ^r	7708	Low level-expression plasmid encoding TALERΔ221L ORF (R144-12aa linker- TS- Δ148)	Section 4.2.1
pJUM500	Km ^r	8913	Over-expression plasmid of TALER6X (R144-6aa linker- TS- Δ148). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM501	Km ^r	8925	Over-expression plasmid of TALER6 (R148-6aa linker- TS- Δ148). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM502	Km ^r	8928	Over-expression plasmid of TALER7 (R148-7aa linker- TS- Δ148). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pJUM503	Km ^r	8934	Over-expression plasmid of TALER9 (R148-9aa linker- TS- Δ148). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM504	Km ^r	8937	Over-expression plasmid of TALER12 (R148-12aa linker- TS- Δ148). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM505	Km ^r	8907	Over-expression plasmid of TALER0 (R148-TS- D148). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM506	Km ^r	8694	Over-expression plasmid of TALERΔ221 (R148-6aa linker- TS- Δ221). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM507	Km ^r	8724	Over-expression plasmid of TALERΔ221 (R148-12aa linker- TS- Δ221). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM508	Km ^r	8892	Over-expression plasmid of TALERΔ153 (R148-0aa linker- TS- Δ153). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM509	Km ^r	9015	Over-expression plasmid of TALERΔ119 (R148-6aa linker- TS- Δ119). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM510	Km ^r	8943	Over-expression plasmid of TALERJRIG (R148- AEAAAKEAAAKA- TS- Δ119). Tn3 NM resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM511	Km ^r	8925	Over-expression plasmid of TALER6-WTR (R148-6aa linker- TS- Δ148). Tn3 wildtype resolvase as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM512	Km ^r	8925	Over-expression plasmid of TALER6-SY (R148-6aa linker- TS- Δ148). Tn3 SY resolvase (G101S D102Y) as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM513	Km ^r	8925	Over-expression plasmid of TALER6-M (R148-6aa linker- TS- Δ148). Tn3 M resolvase (G101S, D102Y, M103I, Q105L) as catalytic domain and TALE1297 as TALE DBD.	Section 4.2.2
pJUM514	Km ^r	8925	Over-expression plasmid of TALER245 (R148-6aa linker- TS- Δ148). Coding sequence of catalytic domain from pJUM045 used, and TALE1297 as TALE DBD.	Section 5.2.1
pJUM602	Km ^r	9027	Over-expression plasmid of TALER100 (R148-6aa linker- TS- Δ148). Tn3 NM resolvase as catalytic domain, and HIV-TALERA as TALE DBD.	Section 5.2.2

Table 2.6- continued

Name	Antibiotic marker	Size (bp)	Description	Source/ Reference
pJUM603	Km ^r	9027	Over-expression plasmid of TALER102 (R148-6aa linker- TS- Δ148). Coding sequence of catalytic domain from pJUM012 used, and HIV-TALERA as TALE DBD.	Section 5.2.2
pJUM604	Km ^r	9027	Over-expression plasmid of TALER104 (R148-6aa linker- TS- Δ148). Coding sequence of catalytic domain from pJUM045 used, and HIV-TALERA as TALE DBD.	Section 5.2.2
pJUM612	Km ^r	9027	Over-expression plasmid of TALER101 (R148-6aa linker- TS- Δ148). Tn3 NM resolvase as catalytic domain, and HIV-TALERB as TALE DBD.	Section 5.2.2
pJUM613	Km ^r	9027	Over-expression plasmid of TALER103 (R148-6aa linker- TS- Δ148). Coding sequence of catalytic domain from pJUM012 used, and HIV-TALERB as TALE DBD.	Section 5.2.2
pJUM614	Km ^r	9027	Over-expression plasmid of TALER105 (R148-6aa linker- TS- Δ148). Coding sequence of catalytic domain from pJUM045 used, and HIV-TALERB as TALE DBD.	Section 5.2.2

2.7 Plasmid Cloning

Plasmids were constructed by inserting annealed oligonucleotides or digested double-stranded DNA fragments into plasmid backbones. Some of the plasmid construction strategies for expression and substrate plasmids are outlined in Section 2.10. Here, the routine laboratory techniques and protocols for plasmid design and amplification are provided.

2.7.1 Annealing oligonucleotides

Oligonucleotides obtained from Eurofins Genomics were re-suspended in an appropriate volume of TE buffer (10 mM Tris-HCl pH 8.2, 1 mM EDTA) to attain a concentration of 100 μ M. Single-strand oligonucleotide pairs were then annealed in a 1:1 ratio at a concentration of 2 μ M in 100 μ l of TE buffer (+ 50 mM NaCl). The components were mixed and heated at 95 °C (or above the oligonucleotides' melting temperatures as indicated on the supplier's synthesis report sheet) for 5 minutes in a heating block. The heating block was turned off and allowed to cool/anneal for \geq 45 minutes or until the block reached a temperature close to 37 °C. Preparation of fluorescently-labelled oligonucleotides for EMSA is described in Section 2.14.2.

2.7.2 Restriction endonuclease digestion of DNA

Plasmids and double-stranded DNA (gBlocks) were digested in NEB's recommended buffer (see below). 5-10 units of restriction enzyme were used per microgram of DNA. The digest reactions were incubated at 37 °C for \geq 1 h and were terminated by the addition of SDS loading buffer before electrophoresis. Restriction digests of *in vitro* CR reaction products (Section 2.16) were done in a similar manner and supplemented with MgCl₂ or pH-adjusted where required.

Buffer	Components
NEBuffer 2.1	50 mM NaCl, 10 mM Tris-HCl, 10 mM MgCl ₂ , 100 µg/ml BSA, pH 7.9@25 °C
NEBuffer 3.1	100 mM NaCl, 50 mM Tris-HCl, 10 mM MgCl ₂ , 100 µg/ml BSA, pH 7.9@25 °C
CutSmart Buffer	50 mM Potassium Acetate, 20 mM Tris-acetate, 10 mM Magnesium Acetate, 100 µg/ml BSA, pH 7.9@25 °C
SDS loading Buffer	50% glycerol, 1% SDS, 0.01% bromophenol blue

2.7.3 Gel electrophoresis (I) - Agarose gel electrophoresis

For separation of digested DNA fragments, the stopped digest mixture was loaded on 0.9-1.0% agarose gels. These gels were made by completely dissolving an appropriate amount of 'Ultrapure' agarose (Invitrogen) in 100 ml of 1×TAE buffer (40 mM Tris-acetate pH 8.2, 1mM EDTA) and heating the mixture in a microwave oven. The hot agarose solution was cooled to ~60 °C, poured into a gel former fitted with an appropriate comb, and then allowed to set at room temperature. Gels were run at room temperature in 1×TAE buffer for 30 to 120 minutes at 100 V depending on fragment size. The sizes of DNA fragments and plasmids were estimated by running NEB 1 kb DNA ladder (500 bp to 10 kb) alongside the samples where required (Fig. 2.1).

DNA products resulting from *in vivo* and *in vitro* recombination reactions were usually separated in a similar manner on 1.2% agarose gels for 120 minutes at 120 V or 16 h at 20 V. Nicked *in vitro* recombination reactions were run on 0.7% agarose gels at 20 V for 16 h (Section 4.2.3).

To visualise DNA on agarose gels, the gel was rinsed in a 0.6 µg/ml ethidium bromide solution (made by diluting a 15 mg/ml ethidium bromide stock solution in 1 x TAE electrophoresis buffer or water) for 30-60 minutes. Background ethidium bromide fluorescence was removed by rinsing the gel several times in water and soaking for a further 30-60 minutes. Ethidium-stained DNA was

visualised by exposing the gel to short wavelength 254 nm UV illumination. Where bands were to be excised from the gels for DNA purification, a long wavelength 365 nm source was used. Gels were photographed using a Gel Doc XR System (Bio-Rad). Experiments reported in this work were generally repeated multiple times ($n \geq 3$), the gel electrophoresis images shown in Chapters 3, 4 and 5 are exemplars.

2.7.4 Extraction of DNA from gel fragments

Upon visualisation (using long wavelength transillumination), bands of interest were excised using a clean scalpel. Excised gel chips were transferred into 0.45 μm filter centrifuge 'Costar' tubes (Corning Inc.) or pre-weighed 'Nunc' tubes (Thermo Fisher Scientific). Where the filter tube was used, the tube was centrifuged at 10 000 rpm for 10 minutes. The filtrate was then collected and transformed directly into cells or first purified by DNA ethanol precipitation. Where Nunc tubes were used, the weight of the gel chip was determined. DNA was then purified from the gel chip using the QIAQuick gel extraction kit (Qiagen) following the manufacturer's instructions (Cat No./ID: 28704). This kit uses a silica-membrane system for DNA purification of 70 bp to 10 kb fragments through rapid bind, wash and elute steps. The purified plasmid DNA was eluted using the provided elution buffer (EB) (10 mM Tris-Cl, pH 8.5). DNA was then used for downstream DNA ligation or stored at $-20\text{ }^{\circ}\text{C}$.

2.7.5 Ligation of DNA restriction fragments

1-2 μg of DNA fragments were ligated in 1 x NEB T4 DNA ligase buffer (10 mM MgCl_2 , 50 mM Tris-HCl, 1 mM ATP, 10 mM DTT, pH 7.5@25 $^{\circ}\text{C}$) using 1 unit of T4 DNA ligase (NEB) in a final volume of 10-20 μl . The molar ratio of insert to vector was usually 3:1. Ligation reactions were carried out at 14 $^{\circ}\text{C}$ overnight and then ethanol-precipitated or used to transform competent *E. coli* cells directly (Section 2.8).

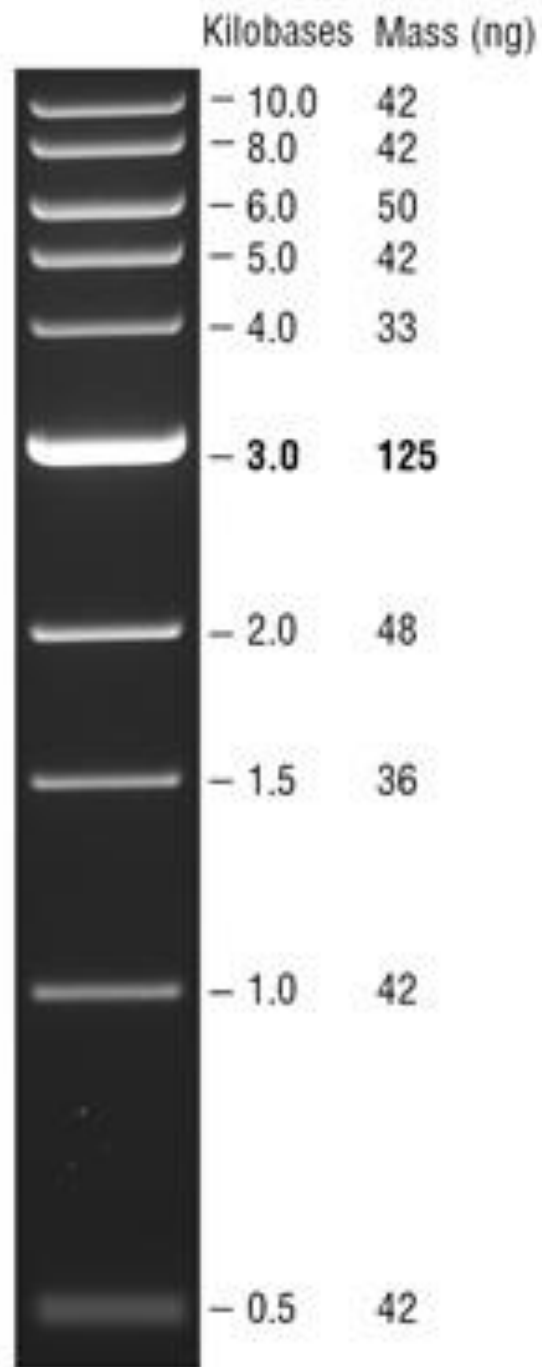


Figure 2.1: 1KB DNA Ladder. This DNA ladder was usually run alongside samples on agarose gels for the estimation of fragment sizes (NEB.com, 2018; Sambrook *et al.*, 1989). The intense 3 kb band is usually a good reference marker when evaluating recombination reaction products.

2.7.6 Ethanol precipitation of DNA

The salt concentration of the DNA solution was adjusted by adding 0.22 volume of 5 M ammonium acetate. 2.5 volumes of 100% ethanol were then added, and the sample was thoroughly mixed and incubated at -20°C overnight. The precipitated DNA was pelleted by centrifugation in a refrigerated Eppendorf microcentrifuge (13 000 rpm, 4°C , 90 min.). The supernatant was carefully removed, and the pellet was washed with 80% ethanol and centrifuged for another 15 min (13 000 rpm, 4°C). The ethanol was removed as before, and the resulting DNA pellet was briefly dried in a Genevac™ Centrifugal Concentrator DNA Integrated System (SP Scientific) and re-suspended in distilled water or TE buffer.

2.7.7 Preparation of competent *E. coli* cells

E. coli cells were prepared for DNA transformation by treatment with CaCl_2 for chemically competent cells or by a series of washes with a 10% glycerol solution for 'electro-competent' cells.

To make chemically competent cells, 200 μl of an overnight culture of DS941 (Section 2.1) was inoculated into 10 ml of L-broth and grown at 37°C with shaking (250 rpm in New Brunswick Scientific Excella E24 Incubator Shaker series) for 90 - 120 mins ($\text{OD}_{600} \approx 0.4-0.5$). Cells were chilled on ice and harvested by centrifugation (Beckman Coulter J-20, 7000 rpm, 4°C , 3 minutes). The pellet was gently re-suspended in 20 ml of ice-cold 50 mM CaCl_2 . Cells were centrifuged as before, gently re-suspended in 10 ml of 50 mM ice-cold CaCl_2 and kept on ice for 30 minutes. The cells were centrifuged again and gently re-suspended in 1 ml of ice-cold CaCl_2 . The cells were now ready to be used for chemical transformation or to be kept at 4°C for up to 48 hours. For short-term storage (1 month) at -70°C , the 50 mM CaCl_2 for the final resuspension was supplemented with 15% glycerol.

For higher transformation efficiencies, electro-competent cells were prepared for long-term storage and use. 400 ml of L-broth (+ antibiotic, if required) was

inoculated with 4 ml of an overnight *E. coli* culture and grown at 37 °C with shaking (300 rpm in New Brunswick Scientific Excella E24 Incubator Shaker series) until OD₆₀₀ of 0.5 - 0.7 was reached. Cells were then chilled on ice to 4 °C for 20 mins before harvesting by centrifugation (Beckman Coulter JA-14, 4000 x g, 4 °C, 15 min). All steps from this stage were done on ice, in the cold room or at 4 °C. After centrifugation, the supernatant was carefully removed, the pellet was gently re-suspended in 400 ml ice-cold 10% glycerol, and centrifugation was repeated. The pellet was re-suspended in 200 ml ice-cold 10% glycerol and centrifuged again. The resulting pellet was then re-suspended in 20 ml of ice-cold 10% glycerol, transferred to a 30-ml polypropylene tube and centrifuged (Beckman Coulter J-20, 4000 x g, 4 °C, 15 min). The final pellet was re-suspended in 1-4 ml of ice-cold 10% glycerol. Aliquots of 200 - 400 µl were then dispensed into Eppendorf tubes and stored at -70 °C.

2.7.8 Transformation of *E. coli* cells

For chemical transformation, 100 µl of competent cells was added to 0.01-0.1 µg of plasmid DNA (or 5 µl of ligation reaction mix) in a chilled Eppendorf tube on ice. The cell-DNA mixture was mixed carefully by gently pipetting up and down and then incubated on ice for 10-30 minutes. The cells were then heat-shocked by incubating the tube at 37 °C for 5 minutes before returning to ice for 2 minutes. 200 µl of L-broth was then added and the cells were incubated at 37 °C for 30 to 90 minutes to allow for expression of antibiotic resistance gene and cell recovery (recovery time is dependent on antibiotic to be used for selection). The cells were then plated out on selective media (contains appropriate antibiotics) and incubated at 37 °C overnight.

E. coli electro-competent cells were transformed using the Bio-Rad MicroPulser™ electroporator. 20 µl of electro-competent cells was added to 0.01-0.1 µg of plasmid DNA on ice and mixed gently before the mixture was transferred to the bottom of a pre-chilled electroporation cuvette (0.2cm gap width). The cuvette was slid into the shocking chamber of the electroporator and an electrical pulse

(MicroPulser™ EC2 setting, $V = 2.5$ kV, 1 pulse) was delivered to the cuvette. The cuvette was removed and immediately, 1 ml of L-broth was added to the cells. The sample was transferred into an Eppendorf tube and incubated at $37\text{ }^{\circ}\text{C}$ for 30 - 90 mins with shaking (250 rpm). The cells were then plated out on selective media and incubated at $37\text{ }^{\circ}\text{C}$ overnight.

2.7.9 Preparation of plasmid DNA

Plasmid DNA was purified from 1 - 5 ml of overnight culture using the QIAprep Spin Miniprep Kit following the manufacturer's instructions (Cat No./ID: 27106). This kit uses a silica-membrane system for DNA purification of 70 bp to 10 kb fragments through rapid bind, wash and elute steps. The purified plasmid DNA was eluted using the provided elution buffer (EB) (10 mM Tris-Cl, pH 8.5). For plasmid purification for *in vitro* reactions, larger volumes of overnight cultures are used and purified at 7 ml per miniprep column.

2.8 Estimating DNA concentration by UV spectrophotometry

The concentrations of plasmid DNA for use in *in vitro* reactions and single-stranded DNA after purification were estimated by measuring absorbance at 260 nm on a Lambda 45 UV/visible spectrophotometer (Perkin Elmer). The concentration was then determined using the approximation that a sample of double-stranded DNA with an absorbance of 1 at 260 nm had a concentration of 50 $\mu\text{g}/\text{ml}$ and that a sample of single-stranded DNA with an absorbance of 1 at 260 nm had a concentration of 33 $\mu\text{g}/\text{ml}$.

2.9 DNA Sequencing

All plasmids used in this work were commercially sequenced by Eurofins Genomics. Sample preparation for sequencing was as specified by Eurofins Genomics.

2.10 Plasmid design and construction

Two main divisions of plasmids were used in this work- expression plasmids for protein expression and substrate plasmids that carry recombination target sites.

2.10.1 Expression plasmids

Expression plasmids are subdivided into two groups based on the level of protein expression. Low-level expression plasmids are used to express chimaeric recombinases for *in vivo* recombination reactions (Section 2.16) while high-level expression (over-expression) plasmids are used for the high-volume expression of chimaeric recombinases for protein purification and subsequent *in vitro* assays.

The low-level expression plasmids allow expression levels suitable for *in vivo* recombination assays and were constructed based on pMS140 (Burke *et al.*, 2004) (Fig. 2.2). pMS140 has a *ColE1* origin of replication, is maintained at ~20 copies per cell and carries an ampicillin resistance marker. It carries a wild-type Tn3 resolvase gene cassette flanked by NdeI and Acc65I restriction sites. Resolvase expression is driven by an unidentified promoter within 400 bp upstream of the NdeI site. Chimaeric recombinase gene variants were swapped within this plasmid backbone using the NdeI/Acc65I restriction sites. In both TALERs and ZFRs, a SpeI restriction site links the resolvase catalytic domain to the DNA-binding domain. This allowed simple modular swapping of domains for the generation of CR variants. A unique EagI site is also present in the conserved GR motif (residues 141 and 142) of the resolvase catalytic domain derivatives of pMS140. This allowed the design of TALER linker length variants by inserting annealed oligonucleotides with EagI/SpeI overhangs into EagI/SpeI digested plasmid backbones. It is important to note that in swapping genes within the TALER architecture, the flexible linker (Section 4.2.2) is usually located before the SpeI site. Swapping the catalytic domain in TALER plasmids was done using the NdeI/EagI restriction sites. Mutant CRs were also generated by strategically designing cassettes containing mutations to replace target regions flanked by unique restriction sites.

Over-expression plasmids are based on pSA1101 (Arnold *et al.*, 1999). pSA1101 has a *ColE1* origin of replication, encodes a kanamycin resistance marker and contains an inducible T7 promoter that drives the expression of the wild-type Tn3 resolvase gene cassette (also flanked by NdeI and Acc65I restriction sites) (Fig. 2.2, Fig. 2.3). CR coding sequences from low-level expression plasmids were swapped into over-expression plasmid backbones using the NdeI/Acc65I restriction sites. To simplify

the purification of ZFRs, a 6 x histidine tag was inserted into the C-terminal end of Zif268 before the stop codon by amplifying the Zif268 gene from a plasmid using overlap extension polymerase chain reaction (PCR). This his-tagged variant was then cloned into a ZFR plasmid. Subsequent cloning of mutant ZFRs for purification only required swapping of the resolvase catalytic domain fragment (NdeI/SpeI) from the low-level expression plasmid into the over-expression plasmid.

2.10.2 Substrate plasmids

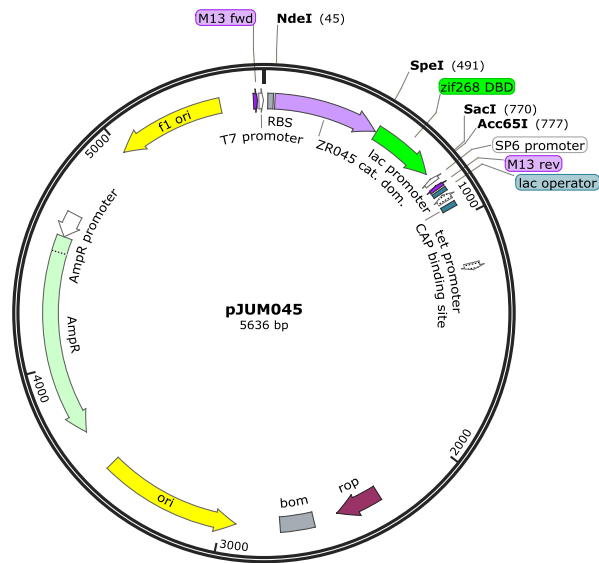
Substrate plasmids carry recombination sites - ZFR target sites called Z-sites and TALER target sites called T-sites. Substrate plasmids contain two identical recombination sites in direct repeat flanking a gene. Two types of substrate plasmids were used in this work- low-copy number plasmids for *in vivo* recombination and high-copy number plasmids for *in vitro* recombination assays. Most of the cloning strategies for both *in vivo* and *in vitro* recombination involved the use of a precursor plasmid, pMTL23 (Chambers *et al.* 1988). pMTL23 carries an ampicillin resistance marker and a deregulated *ColE1* origin of replication. It has a very high copy number making it very useful as a cloning vector. Single recombination Z-site or T-site pMTL23-based plasmids were cloned by inserting annealed oligonucleotides with EcoRI/SacI overhangs into the digested pMTL23 plasmid backbone.

Recombination substrate plasmids used for *in vivo* assays in this study have the two recombination sites flanking a galactokinase gene (*galk*) (Section 2.11) and are based on a pMS183 Δ backbone. pMS183 Δ has a pSC101 replication origin and carries a kanamycin resistance marker (Prorocic, 2009). Two different approaches were taken to generate these two-site substrate plasmids. The first involves the use of the single-site precursor pMTL23 plasmid; a four-piece ligation of BglII/NcoI and BsrGI/XbaI fragments from pMS183 Δ to recombination site-containing XbaI/NcoI and BamHI/Acc65I fragments. This cloning is made possible because of compatible ends generated from restriction by enzymes BsrGI and Acc65I, and BglII and BamHI respectively (Fig. 2.4). The second approach eliminates the use of the precursor plasmid and allows the direct cloning of two pairs of annealed oligonucleotides with different compatible ends directly into pFM141, a two-site

pMS183 Δ -based plasmid. HIV Z-substrate plasmids, pJU001, pJU002 and pJU003, were generated using this approach (Figure 2.5).

Recombination substrate plasmids used for *in vitro* assays in this study have the two recombination sites flanking a kanamycin resistance gene (Section 2.16) and are based on the pMTL23 backbone (Fig. 2.6). These substrate plasmids were constructed by a three-piece ligation of AlwNI/BglII and AlwNI/BamHI recombination site-containing fragments from a pMTL23-based single site plasmid, and a kanamycin resistance-encoding BamHI/BamHI fragment from pUC71K. The orientation of the inserted fragment in the plasmids generated was determined by a NruI digest. The orientations of the selected plasmids are indicated against each substrate plasmid in Table 2.6. Substrate plasmids for *in vitro* assays were also cloned by inserting annealed oligonucleotides with different compatible ends in a one-step four-piece ligation into a two-site pMTL23-based plasmid, pFM160, in the exact approach described in Figure 2.5.

A Low-level expression plasmid



B Over-expression plasmid

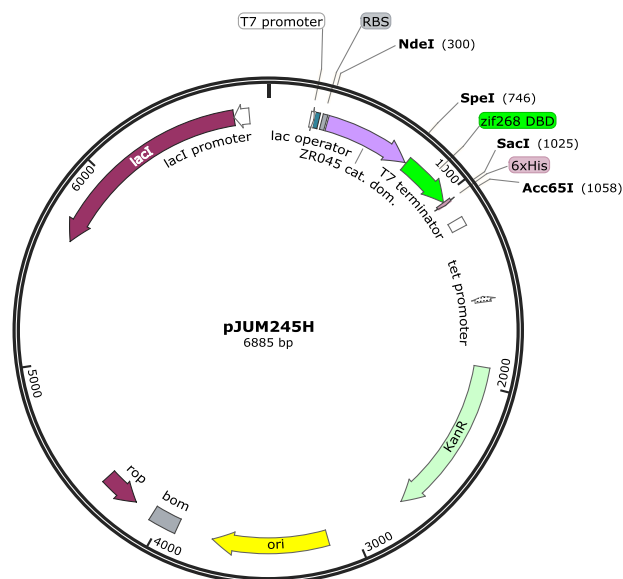


Figure 2.2: Expression plasmid maps using ZR045 as an example. A. ZR045 expression from the low-level expression plasmid is constitutively on. The plasmids are present at a low copy number in individual cells (Section 3.3.8). **B.** With the over-expression plasmid, ZR045 expression is regulated by the LacI repressor and the induction of gene expression is by the addition of IPTG. The antibiotic resistance gene in both plasmid types are different. After the characterization of ZR045, it was cloned from pJUM045 into a his-tagged ZFR overexpression plasmid, pJUM212H using the NdeI/SpeI restriction sites. This cloning resulted in pJUM245, the overexpression plasmid of ZR045 (5.2.1).

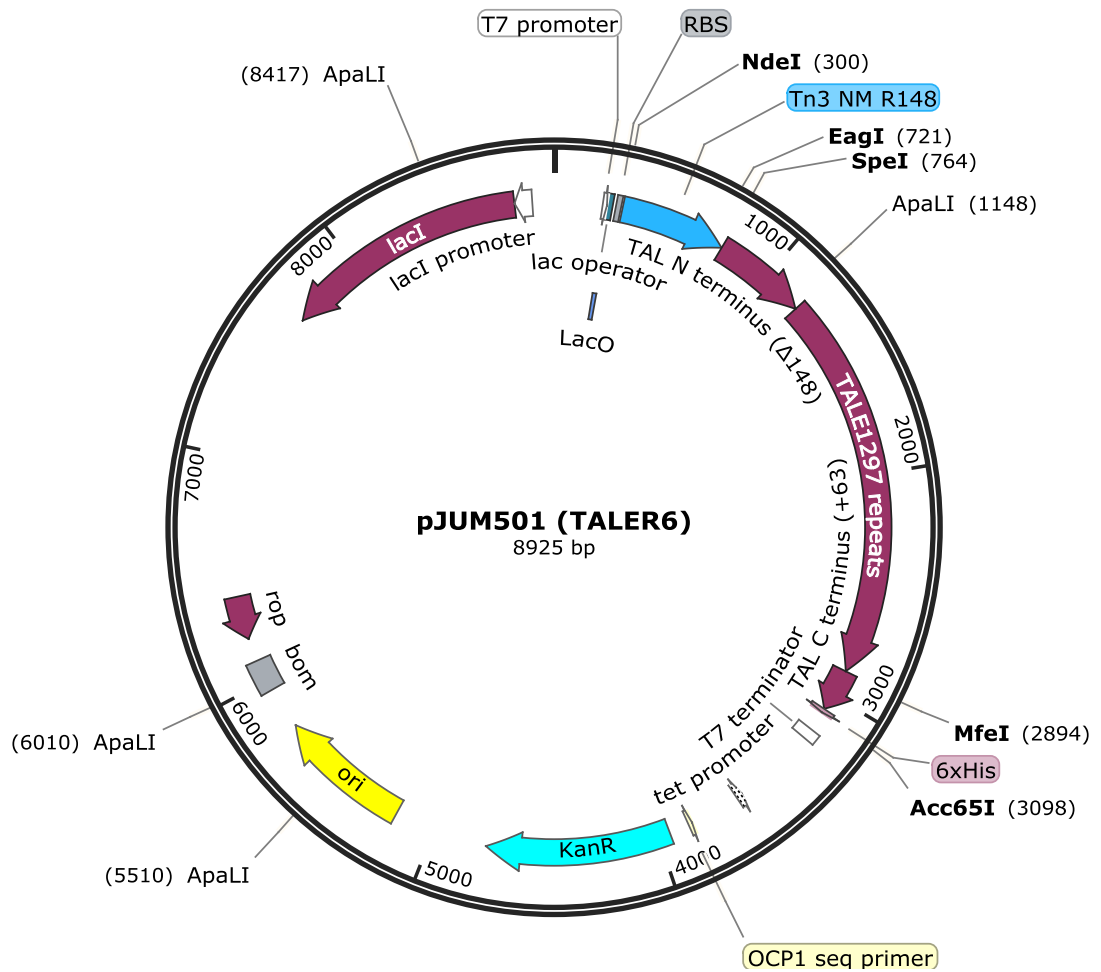


Figure 2.3: Over-expression plasmid map of TALERs using TALER6 as an example. The design of TALER overexpression plasmids is similar to that of ZFRs. Unique restriction sites were used to generate linker length and catalytic domain variants of TALER overexpression plasmids. Linkers of varying lengths were cloned as annealed oligonucleotides into EagI/SpeI digested backbones of TALERs. Catalytic domain variants were swapped between plasmids using the NdeI/EagI restriction sites.

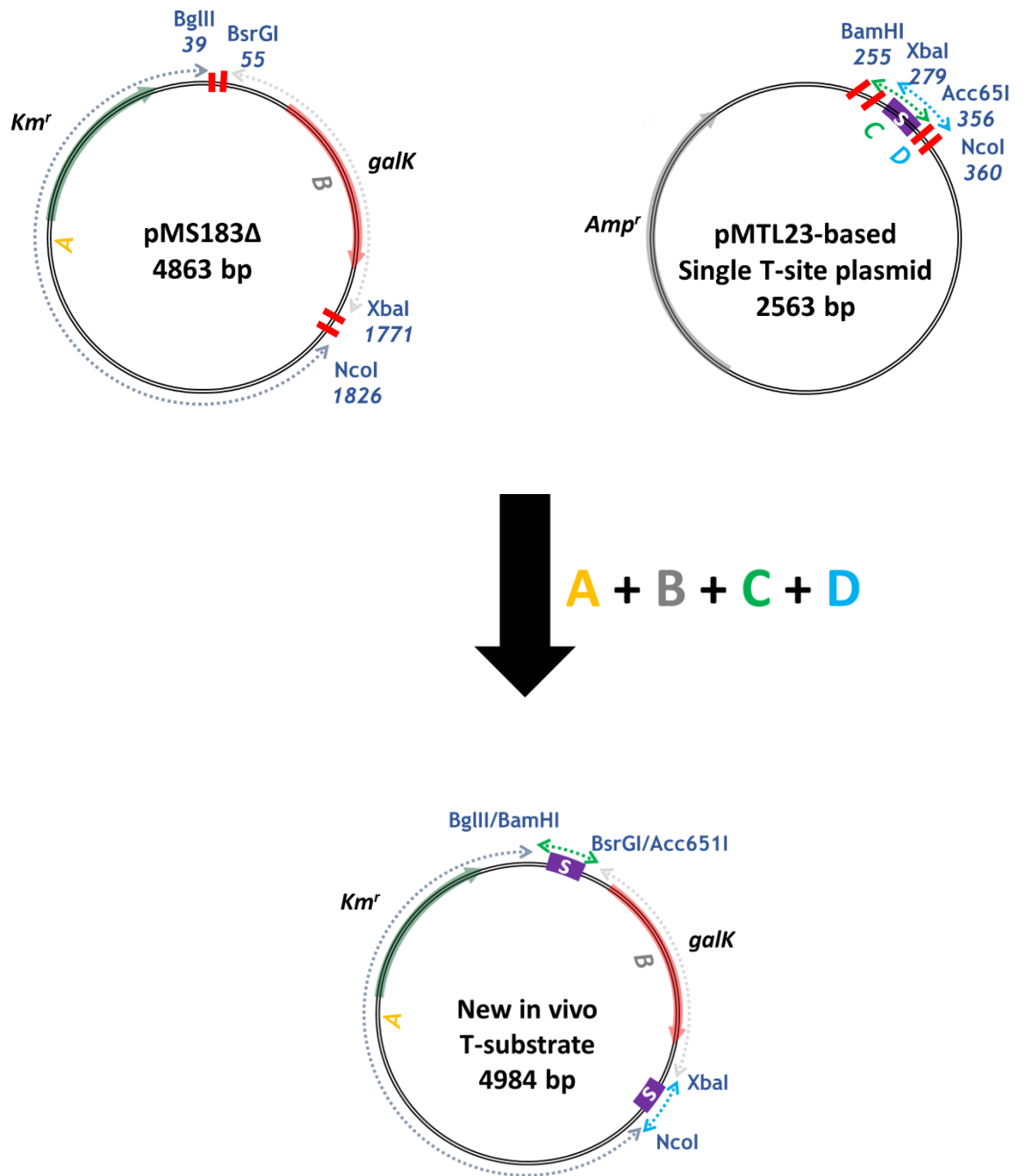


Figure 2.4: Schematic illustrating the construction of new *in vivo* TALER recombination substrate plasmids using a precursor plasmid. The positions of the kanamycin resistance gene and the galactokinase gene are indicated using red and green arrows respectively. The T-sites are shown as purple boxes. Annealed oligonucleotides with EcoRI/SacI sticky ends carrying the T-sites were first cloned into the digested pMTL23 plasmid. They were then cloned into *galk*-expressing pMS183Δ on BamHI/Acc65I and XbaI/NcoI fragments using a four-piece ligation protocol. The final T-substrate plasmids have two T-sites flanking the *galk* gene in direct repeat.

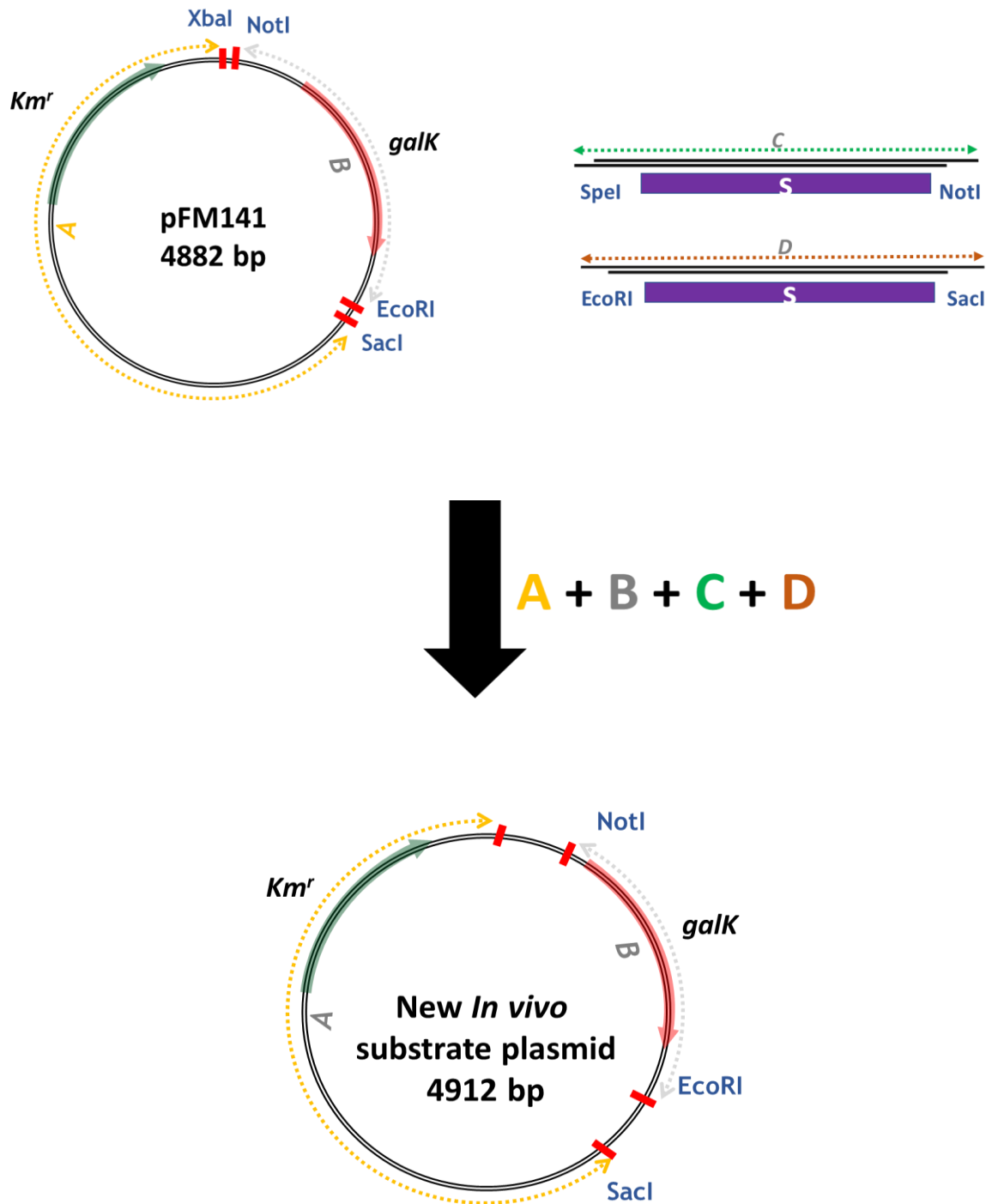


Figure 2.5: Schematic illustrating the construction of new *in vivo* recombination substrate plasmids without using a precursor plasmid. The positions of the kanamycin resistance gene and the galactokinase gene are indicated using red and green arrows respectively. The Z-sites are shown as purple boxes. Here, annealed oligonucleotides with SpeI/NotI and EcoRI/SacI sticky ends carrying the T-sites were cloned directly into digested *galK*-expressing pFM141 using a four-piece ligation protocol or sequentially. The final T-substrate plasmids have two recombination sites flanking the *galK* gene in direct repeat. HIV Z-substrate plasmids were cloned using this strategy.

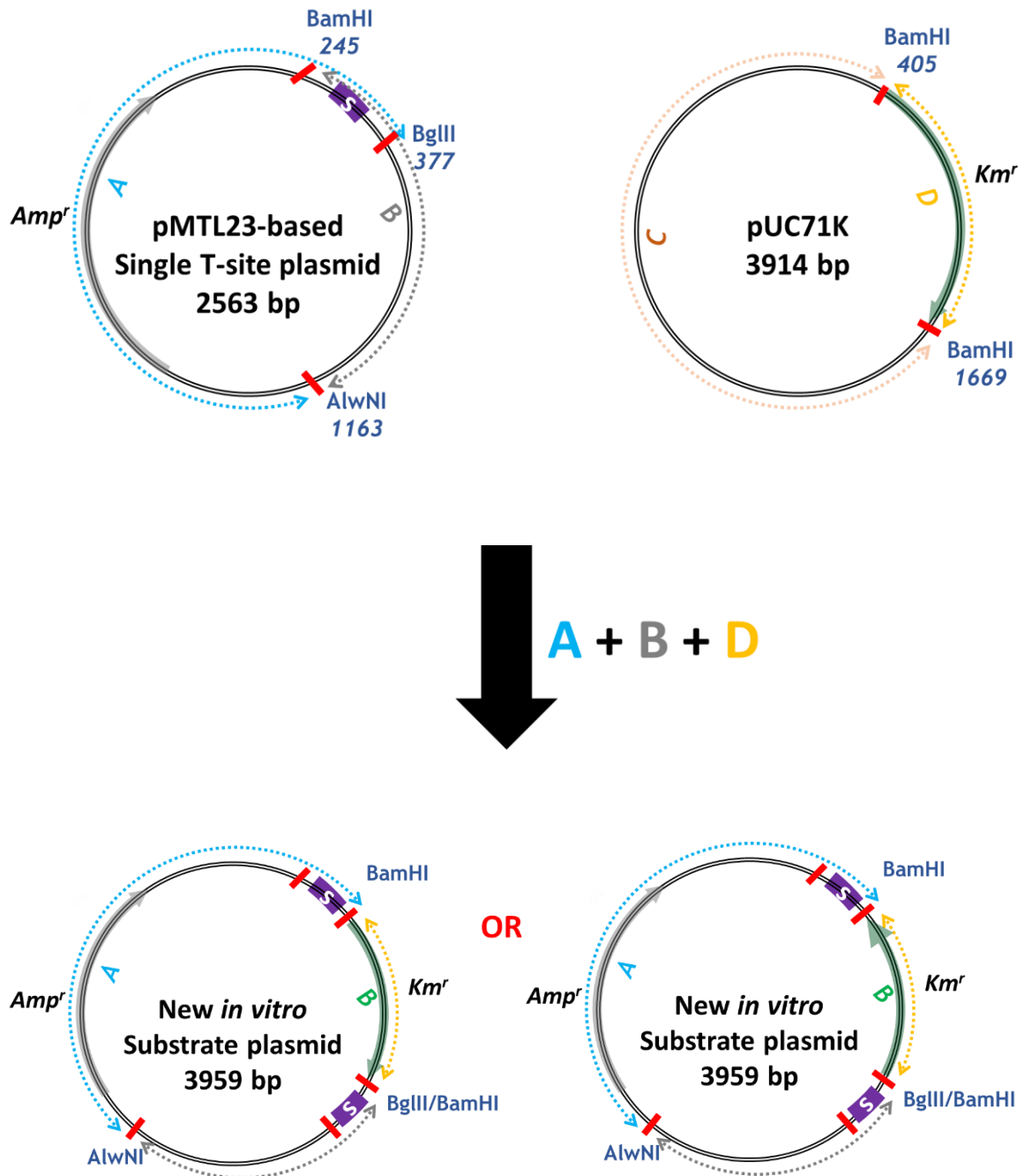


Figure 2.6: Schematic illustrating the construction of new *in vitro* recombination substrate plasmids using a precursor plasmid. The position of the kanamycin resistance gene is indicated using a green arrow. The recombination sites (s) are shown as purple boxes. BamHI/AlwNI and BglIII/AlwNI fragments from the pMTL23-based precursor plasmid were ligated with the BamHI/BamHI kanamycin cassette from pUC71K. A NruI diagnostic digest is used to determine the orientation of the Kanamycin cassette. The final substrate plasmids have two recombination sites flanking the *galk* gene in direct repeat.

2.11 *In vivo* recombination reactions - “MacConkey agar assay”

The *in vivo* recombination assay is based on the excision of the *galk* gene by an active CR due to recognition of the flanking Z-sites or T-sites, and this is reported by the ability or inability of the cell to ferment galactose (Blake, 1993) (Fig. 2.7). The culture medium used, MacConkey agar base (Difco™, BD) is selective for gram-negative bacteria and contains no added carbohydrates. It contains bile salts, peptones and neutral red dye. Neutral red is pH-indicative, changing from red to yellow as pH changes from 6.8 to 8.0. When supplemented with 1% galactose, this agar is indicative of galactose fermentation by giving red colonies.

On these plates, colonies that ferment galactose appear red with or without surrounding bile precipitate due to pH decrease with acid production. Non-fermenting colonies appear colourless due to the utilization of amino acids in peptones as carbon source, resulting in a pH increase that changes the local indicator dye colour. The *E. coli* strain used for this assay, DS941, is a *galk*⁻ strain where the galactokinase gene, *galk*, important for galactose utilization has been knocked out. The presence of a *galk*-containing substrate plasmid in the cells restores galactose utilization. The activity of a recombination-proficient CR on the target substrate plasmid leads to the formation of two circular DNA molecules—one containing the kanamycin resistance gene and the origin of replication, and the other carrying the *galk* gene. The *galk*-harbouring molecule has no origin of replication and is diluted out as the cells multiply.

E. coli cells containing the low copy number substrate plasmid to be tested (Section 2.10.2) were made competent (Section 2.7.7) and transformed (Section 2.7.8) with ~0.1 µg of low-level expression plasmid. After the transformation expression stage, cells were plated out on MacConkey agar plates containing 0.1% galactose, and kanamycin and ampicillin at selective concentrations (Section 2.4). The plates were incubated overnight at 37°C.

To confirm the results of the MacConkey assay, the DNA from the cells was analysed using one of four approaches. Second-day growth of cells as in approach 1 typically yields inconclusive results. Cells recovered from MacConkey agar plates as in approach 2 do not routinely give good recovery of plasmid DNA. Approaches 3 and 4 yielded the most consistent results and were most preferred in this work.

1. 1 ml of L-broth is added to the MacConkey agar plate after the overnight incubation. The cells were then scraped, and the mixture was transferred into an Eppendorf tube and vortexed briefly to break the colonies up and mix thoroughly. 50 μ l of the cell mixture was then transferred into a test tube containing 5 ml L-broth (+ kanamycin) and grown overnight. Plasmid DNA was isolated from pelleted cells (Section 2.7.9) and analysed by agarose gel electrophoresis (Section 2.7.3).
2. 1 ml of L-broth was added to the MacConkey agar plate after the overnight incubation. The cells were then scraped off, and the mixture was transferred into an Eppendorf tube and vortexed briefly to break the colonies up and mix thoroughly. DNA was prepared from 300 μ l of the cell mixture directly (Section 2.7.9) and analysed by agarose gel electrophoresis (Section 2.7.3).
3. After the transformation expression stage, the same volume of cells plated on the MacConkey agar plate was concurrently spread on L-agar plates containing kanamycin and ampicillin. After overnight incubation, DNA was prepared from the cells on L-agar plates as in 2 above.
4. After the transformation expression stage, 50 μ l of cells was incubated in 5 ml of L-broth (containing kanamycin and ampicillin) or similar 1 in 100 dilutions. After overnight incubation, DNA was prepared from the cells as in Section 2.7.9.

DNA prepared was run agarose gels and photographed as described in Section 2.7.3. Where required (e.g. library screening or DNA product analysis), expression plasmids and resolution products were excised and purified from agarose gels as described in Section 2.7.4. Bulk sequencing was carried out by sequencing gel-purified resolution products directly (Section 2.9).

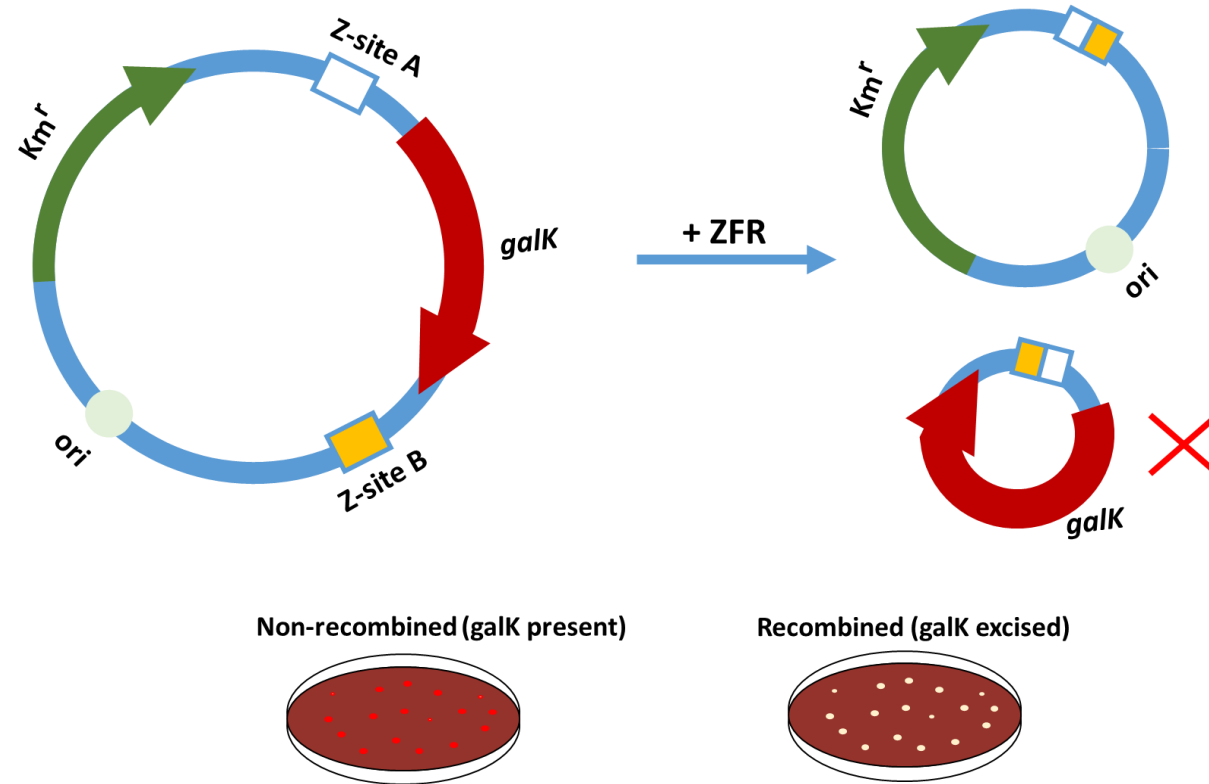


Figure 2.7: *In vivo* recombination assay. An active ZFR catalyses the excision of the *galk* gene, resulting in the loss of ability to utilize galactose as a carbon source. Colonies harbouring recombined plasmids appear 'white' (pale-coloured), while colonies carrying plasmids that still have the *galk* gene appear red on the agar plate.

2.12 Protein Expression and Purification

2.12.1 Large scale induction of CR variants

Chemically competent *E. coli* BL21 (DE3) pLysS cells were transformed with CR over-expression plasmids and grown overnight on selective plates (Section 2.7.8). Single bacterial colonies were picked from the plates and used to inoculate 10 ml selective starter cultures which were incubated overnight at 37 °C. 8 ml of starter cultures were inoculated into 400 ml of warm (37°C) L-broth (+Kan 50 µg/ml and Cm 25 µg/ml) in large conical flasks. The cells were grown in a flat-bed shaker incubator at 37°C with shaking at 250 rpm to an optical density within the range 0.5–0.6. The flasks were cooled down rapidly (in the cold room or on an ice bucket) to 20 °C. A 100 µl sample was collected, pelleted and frozen for future analysis here. 2 ml of 100 mM IPTG (final concentration 0.5 mM) was then added to the flasks to induce protein expression. The cells were returned to the shaker incubator and grown for a further 3 hours. Another 100 µl sample was collected here as before. To harvest the cells, the cultures were spun at 9000 rpm for 10 minutes (4 °C). The pellets were then washed gently with 200 ml of cell wash buffer (20 mM Tris.HCl, pH 7.5 + 10 mM MgCl₂). The cells were harvested by centrifugation as before. The cell pellets weighed approximately between 2.5 to 3 g. The cells were then taken to the next step of protein purification or frozen at -20 °C.

2.12.2 Extraction and purification of His-tagged CRs

The extraction and purification protocols for His-tagged ZFRs and TALERs were quite similar with minor differences in the buffers used. The buffers for CR protein purification are listed in Table 2.7 and Table 2.8. The procedure for the extraction and purification of his-tagged TALER proteins is described below.

If previously frozen, cell pellets were thawed on the bench, and transferred back to ice right after thawing. The cells were re-suspended in 25 ml Lysis buffer on ice using a pipette and then transferred to a chilled 250-ml glass beaker for

sonication. The cells were sonicated using a Vibra-cell VC100 sonicator with a micro-probe at 40% amplitude for 20 seconds three times. 300 µl of PMSF was added immediately after the first sonication to prevent CR proteolysis. The cells were cooled in an ice-water slurry with gentle swirling for 2-5 min after each sonication step to prevent overheating. The crude extracts were transferred to a Dounce homogeniser and homogenised (50 strokes) for 30 minutes to encourage CR solubilization. A 10 µl sample was collected here and frozen. The samples were then transferred to a 30-ml centrifuge tube and spun at 18 500 rpm (Beckman Coulter, JA-20 rotor) for 30 min at 4 °C. The supernatant was collected in a 50 ml Corning centrifuge tube and kept on ice. The pellet was labelled appropriately and stored away at - 20 °C. Another 10 µl sample was collected frozen here.

The CRs used in this work have a C-terminal hexahistidine tag that allowed their purification from contaminating non-CR protein, nucleic acids and cell components using a 1 ml nickel-charged immobilized metal affinity chromatography (IMAC) column (HisTrap HP 5 x 1ml column, GE Life Sciences). The CR protein binds to the column, whilst the unbound contaminating components are washed away. The CR protein is then eluted with a sufficient concentration of imidazole which competes with the His-tag for the nickel resin. All chromatography steps were carried out at room temperature on an AKTA Purifier HPLC system (Amersham Biosciences).

The HPLC system was first prepared by washing the lines with distilled water with absorbance readings set to 280 nm, 260 nm, and 213 nm for measuring absorbance during the purification run. The lines were then filled with the appropriate buffers - TALER Gradient Buffer A and TALER Gradient Buffer B (Table 2.7). The 1-ml HisTrap HP column was connected to the pump running at 1 ml/min. 5 ml of distilled water was used to wash out its preservation buffer and then the column was prepared for sample loading by passing 10 ml of Gradient Buffer A through it. After this, the column was loaded with the CR-containing supernatant from the 50 ml Corning centrifuge tube at a flow rate of 1 ml/minute. The column was then re-equilibrated with 30 ml of TALER Gradient Buffer A to wash off unbound contaminants. The CR protein was eluted by running a gradient to 100% of TALER

Gradient Buffer B against TALER Gradient Buffer A over 25 minutes. 1-ml fractions of purified CR protein were collected in Nunc tubes in the fraction collector. Selected fractions determined by A_{280} peaks were labelled and kept on ice. Usually two peaks were recorded, the second being a peak representing the purified CR. Fractions corresponding to absorbance peaks were run on a discontinuous polyacrylamide gel to determine purity (Fig. 2.8).

Selected individual fractions or multiple fractions pooled together were bagged and dialyzed overnight in Glycerol Storage Dialysis Buffer. Purified proteins were harvested, transferred into labelled Nunc tubes and stored at $-20\text{ }^{\circ}\text{C}$.

Minor variations to the TALER purification protocol were applied in the his-tagged ZFR purification. 6 M Urea was incorporated into some of the buffers for denaturation (Table 2.8). Cell pellets were thawed and sonicated in a low-salt concentration ZFR Lysis buffer. After sonication, the sample was transferred to a 30-ml centrifuge tube and spun at 18 500 rpm for 15 min at $4\text{ }^{\circ}\text{C}$. The supernatant was collected in a glass bottle and stored away at $-20\text{ }^{\circ}\text{C}$. The pellet was then re-suspended in 25 ml of lysis buffer, homogenised in the Dounce homogeniser for 10 minutes and spun down again as before with the supernatant collected and stored away. The pellet was then re-suspended in the ZFR Solubilization buffer and homogenized on ice for 30 minutes. After this, the sample was spun down at 18 500 rpm for 25 min at $4\text{ }^{\circ}\text{C}$. The supernatant collected here was loaded on the AKTA purifier as described for TALER proteins. The ZFR Imidazole-Urea Buffer A and ZFR Imidazole-Urea Buffer B differ from the TALER gradient buffers in that they contain urea. His-tagged ZFR protein fractions were dialysed directly into Glycerol Storage Dialysis buffer.

Table 2.7: Buffers used in the purification of His-tagged TALER variants

Buffer	Composition
Suspension/Lysis buffer	50 mM sodium phosphate pH 7.2, 1000 mM NaCl, 1 mM DTT, 1 mM PMSF, 50 mM Imidazole
Gradient Buffer A	50 mM sodium phosphate pH 7.2, 1000 mM NaCl, 1 mM DTT, 50 mM Imidazole
Gradient Buffer B	50 mM sodium phosphate pH 7.2, 1000 mM NaCl, 1 mM DTT, 500 mM Imidazole
Glycerol Dialysis buffer	25 mM Tris-HCl pH 7.5, 1 mM DTT, 1000 mM NaCl, 50% glycerol.
TALER Dilution buffer	25 mM Tris-HCl pH 7.5, 1 mM DTT, 1000 mM NaCl, 50% glycerol.

Table 2.8: Buffers used in the purification of His-tagged ZFR mutants

Buffer	Composition
Lysis buffer	50 mM sodium phosphate pH 7.2, 200 mM NaCl, 1 mM DTT, 1 mM PMSF
Solubilization buffer	50 mM sodium phosphate pH 7.2, 1000 mM NaCl, 1 mM DTT, 50 mM Imidazole, 6M Urea
Imidazole-Urea Buffer A	50 mM sodium phosphate pH 7.2, 1000 mM NaCl, 1 mM DTT, 50 mM Imidazole, 6M Urea
Imidazole-Urea Buffer B	50 mM sodium phosphate pH 7.2, 1000 mM NaCl, 1 mM DTT, 500 mM Imidazole, 6M Urea
Glycerol Dialysis buffer	25 mM Tris-HCl pH 7.5, 1 mM DTT, 1000 mM NaCl, 50% glycerol.
ZFR Dilution buffer	25 mM Tris-HCl pH 7.5, 1 mM DTT, 1000 mM NaCl, 50% glycerol.

2.13 Estimating protein concentration

The concentrations of purified CR samples were estimated in this work using three different approaches: A_{280} extinction measurements, denaturing gel electrophoresis, and the Bradford assay. None of these approaches have been deemed ideal and concentrations used in this work should be treated as optimum estimations and not precise measurements.

- A_{280} extinction measurements: In this approach, the measured absorbance of a CR protein dilution at OD_{280} is related to its concentration based on its molar extinction coefficient. This extinction coefficient is dependent on its tryptophan (W), tyrosine (Y) and cysteine (C) amino acid composition. As these residues were very minimally available in both ZFRs and TALERS,

concentrations determined using this approach were not considered fit for use.

- Denaturing gel electrophoresis: Concentrations of purified CR samples were estimated by comparing their dilutions with protein standards of known concentrations on a discontinuous SDS-polyacrylamide gel (Section 2.14.1). The protein fractions used as standards in this study were standardised based upon the amino acid analysis performed by D. Blake (Ph. D. Thesis, 1993) on the wild-type Tn3 resolvase fraction R17 f.47 (W.M. Stark & F.J. Olorunniji, personal communication). These proteins were Tn3 NM ZFR (estimated concentration, 40 μM), Tn3 NM resolvase (estimated concentration, 100 μM) and ΦC31 fusion protein (estimated concentration, 16 μM). This approach was error-prone as visual inspection of serially-diluted pure CR fractions was used to determine relative concentrations based on the standards (Fig. 2.9).
- Bradford assay: The Bradford assay was used to determine the concentration of TALER proteins. As the parallel comparison of the activities of these proteins was being made, it was essential to obtain equitable concentrations for analysis. The assay was carried out using the Bio Rad protein assay dye reagent concentrate (Catalog number: 500-0006) and following the manufacturer's protocol. The assay yields a colorimetric response based on the interaction of its dye constituent, Coomassie brilliant blue G-250, to basic and aromatic amino acid residues in the sample protein. Spectrophotometric measurements were made to obtain values which allow the deduction of protein concentrations by comparing OD_{595} for sample proteins to those generated for bovine serum albumin (BSA) of known concentrations on a standard curve. It is important to note that the Bradford assay can give fairly unreliable measurements; however, care was taken here to estimate concentrations of compared proteins under very similar conditions.

2.14 Gel electrophoresis (II) - Polyacrylamide gel electrophoresis

Polyacrylamide gels were used in this work for the analysis of purified protein fractions, the characterization of protein binding activity on double-stranded DNA fragments and for purifying oligonucleotides. Varying concentrations of acrylamide, different types of running buffers, with or without the addition of similar sodium dodecyl sulfate (SDS) were used in each case but gels were always constructed in a similar manner.

Clean glass plates were clamped together with 0.75 mm spacers between them with one of the plates wrapped around the edges with rubber tubing to form a seal. The acrylamide gel mixture was poured between the plates and a well-forming comb was immediately inserted at the top of the plates. The acrylamide gel mix was then left to polymerise for ≥ 1 h before the clamps, tubing and comb were removed. The sealed glass plate containing polymerised gel was then clamped into the electrophoresis kit. The buffer reservoirs were filled with the appropriate running buffers and the gels were run.

2.14.1 Discontinuous polyacrylamide gel electrophoresis

Purified protein fractions and samples collected during the purification steps were analysed using a discontinuous sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) system (Laemmli, 1970). This system involves running the samples through two gels of different acrylamide percentages and different ionic strengths. First, the sample passes through a low-acrylamide percentage low-ionic strength stacking gel before entering the higher-acrylamide percentage higher-ionic strength resolving gel.

The resolving gel used was typically prepared from a solution with a following composition: 10% acrylamide premix (37.5:1 ratio acrylamide to bisacrylamide), 375 mM Tris-HCl (pH 8.8), 0.1% SDS, 0.1% ammonium persulphate (APS), 0.05% (v/v) tetraacetylenediamine (TEMED). For ZFRs, the acrylamide composition of the resolving gel was 15%. This resolving gel was poured between two sealed

glass plates, overlaid with isopropanol to exclude air, and allowed to polymerise for 30-45 minutes. After polymerisation, the isopropanol was removed by rinsing the gel surface with distilled water and blotting off the excess water. The stacking gel was then poured on the polymerised resolvasse gel. The stacking gel is composed of: 5% acrylamide premix (37.5:1 ratio acrylamide to bisacrylamide), 125 mM Tris-HCl (pH 6.8), 0.1% SDS, 0.1% APS, 0.2% (v/v) TEMED. The gel comb was placed in immediately and the stacking gel was allowed to polymerise for 30-45 minutes. After polymerisation, the comb was removed, and the formed wells were rinsed with the Electrophoresis running buffer (Table 2.9). Prior to loading, the Laemmli loading buffer (Table 2.9) was added to protein samples (20% of the total volume) and the samples were boiled for 5 minutes to help denature the proteins and reduce disulphide bonds. Samples were loaded and polyacrylamide gels were run at 200 V for 3 hours.

Table 2.9: Buffers for discontinuous SDS-PAGE

Buffer	Composition
Electrophoresis running Buffer (Tris-glycine buffer)	25 mM Tris base, 250 mM glycine, 0.1% SDS
Laemmli loading Buffer	50% glycerol, 5% SDS, 200 mM Tris-HCl pH 6.8, 0.1 mM EDTA

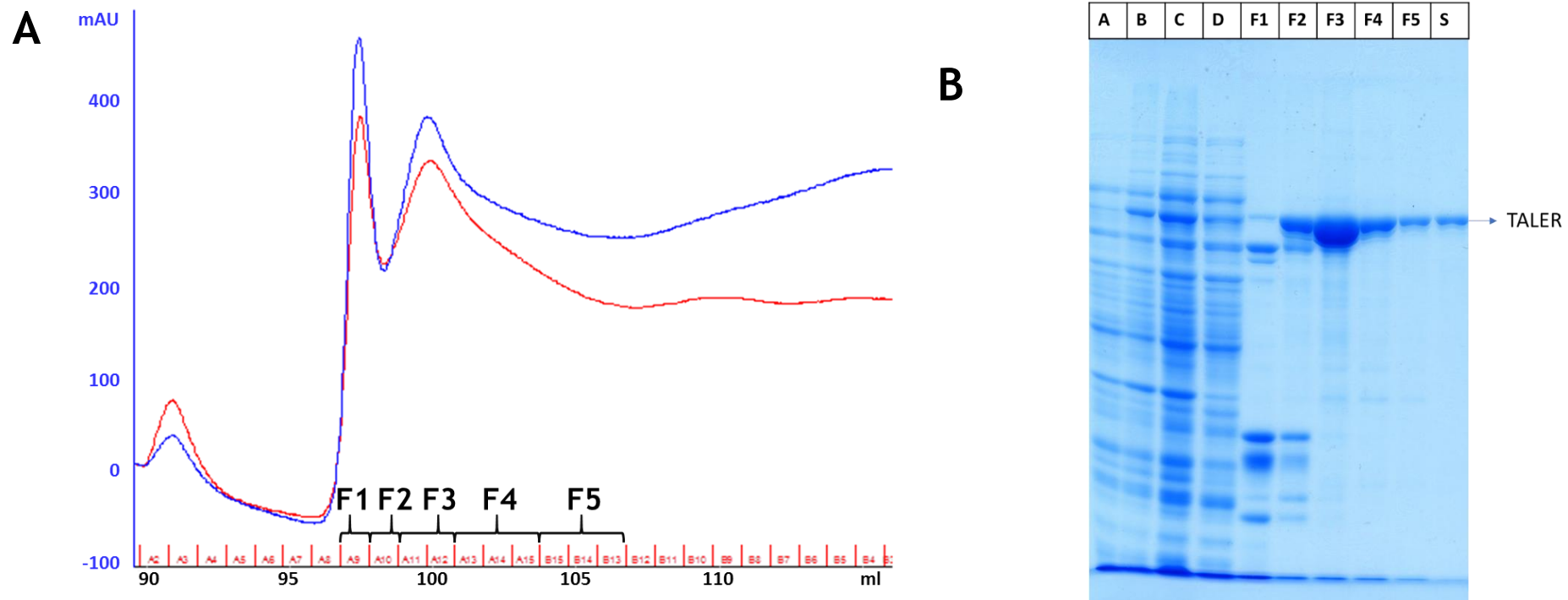


Figure 2.8: Analysis of TALER245 protein purification fractions. **A.** Akta Purifier chromatogram output showing A_{280} (blue) and A_{260} (red) trace at the elution stage of the purification of TALER245. Two major peaks are usually observed with TALER purification protocols with the target protein at the second peak. Absorbance measurements are provided in milli-Absorbance Units (mAU). Selected fractions were dialysed individually or pooled together. **B.** Discontinuous SDS-PAGE gel for determining protein purity. Samples from four fractions collected at different stages of the protein purification process along with samples from dialysed purified TALER245 fractions were run. The abbreviations used are as follows 'A' (sample taken prior to induction), 'B' (sample after 3-hour induction), 'C' (sample collected after homogenization), 'D' (sample loaded on column), 'F1' ... 'F5' (pooled fractions 1 to 5), 'S' (8 μ M TALER6 standard control).

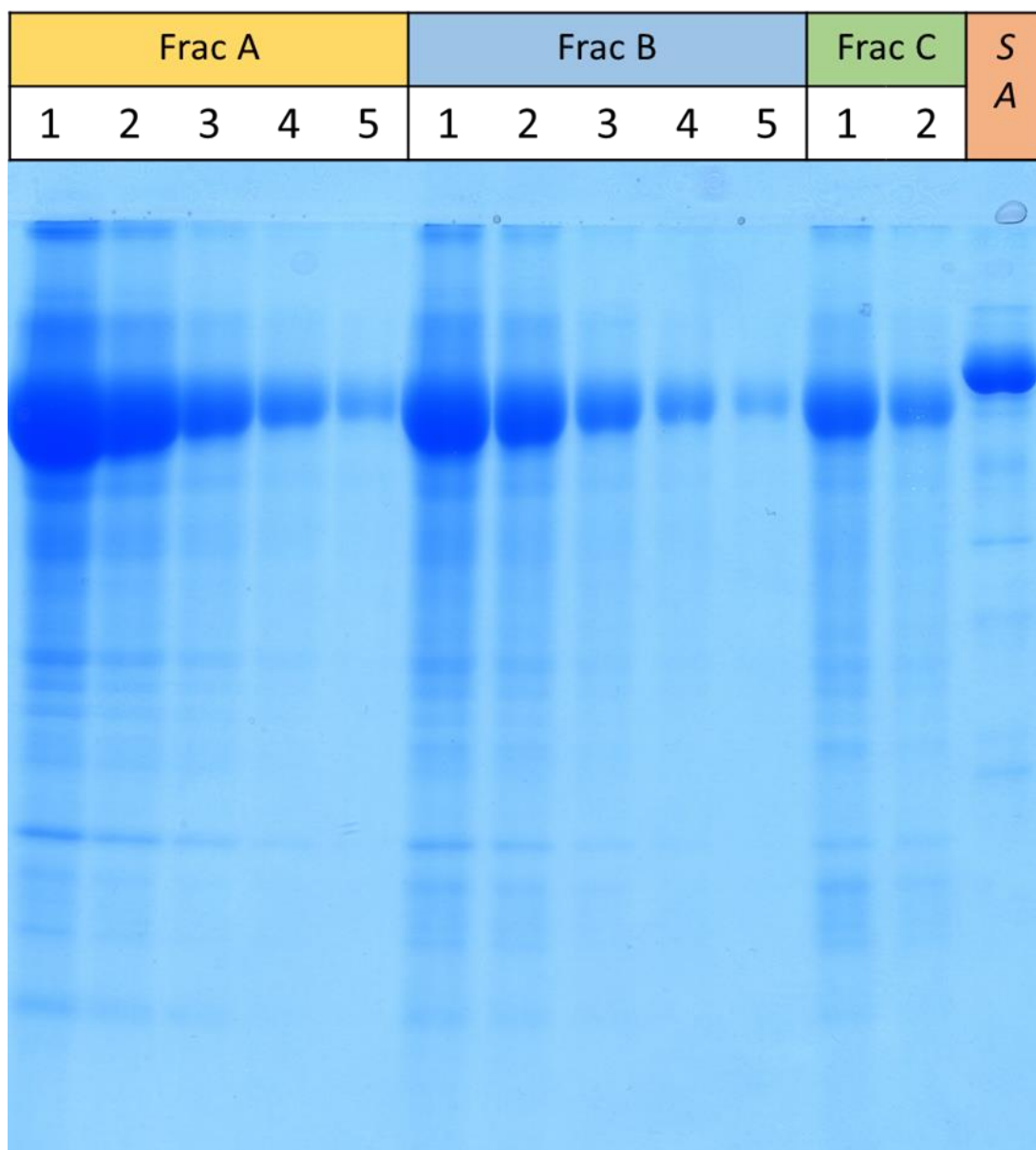


Figure 2.9: TALER6X concentration determination using SDS-PAGE. Selected fractions of the dialysed protein (Frac A, Frac B and Frac C) were serially diluted and run on a discontinuous SDS-PAGE gel alongside proteins of known concentration. Protein concentration was estimated by comparing band intensity of test protein with standards. Dilutions loaded were 1 (stock), 2 (1/2), 3 (1/4), 4 (1/8) and 5 (1/16). The standard protein used was SA (Φ C31 fusion protein with estimated concentration, 16 μ M).

2.14.2 Denaturing PAGE for oligonucleotide purification

After obtaining synthetic oligonucleotides required for the analysis of TALER binding activity from Eurofins genomics, desired full-length oligonucleotides were purified by excision from a SDS-PAGE gel. Synthetic single-strand oligonucleotides were run on 15% polyacrylamide/7 M urea denaturing gels which were made by mixing 15.75 g of urea, 14.75 ml of a 40% (w/v) acrylamide solution (19: 1 ratio of acrylamide to bisacrylamide), 3.75 ml of 10 x TBE, 7.75 ml of H₂O, 225 µl of 10% APS (w/v) and 18 µl TEMED. The electrophoresis running buffer used here was 1 X TBE (Table 2.10). The unloaded gel was first run at 400-500 V (constant voltage) for 30 minutes to heat the gel up before sample loading.

Samples were mixed with formamide loading buffer (Table 2.10) in a 1:1 ratio before they were heated to 95 °C for 5 minutes. The samples were then loaded onto the gel which was run at 400-500 V (constant voltage) for 90 minutes to keep the gel hot, preventing the renaturation of single-stranded oligonucleotides. The gel was stained with 'Stains-all' (1-ethyl-2-[3-(1-ethylnaphtho[1,2-]thiazolin-2-ylidene)-2-methylpropenyl)naphtho[1,2-d]thiazolium bromide; from Aldrich). 70 ml of water and 20 ml of isopropanol were mixed with 10 ml of a 0.1% (w/v) solution of 'Stains-all' in formamide.

The gel was swirled in this solution for 5 minutes or until purple bands appeared, and then de-stained by rinsing with water several times. The desired full-size oligonucleotides to be purified were excised from the acrylamide gel using a scalpel and then transferred to a Nunc tube. The gel chips were crushed with clean pestles and 750 µl of TE buffer was added to the tubes. The mixture was then incubated on a thermomixer (250 rpm) at 37 °C overnight. To recover the DNA, the gel/TE slurry mixture were centrifuged through 0.22 µm filter centrifuge 'Costar' tubes (Corning Inc.) at 13 000 rpm for 1 minute. The supernatant was transferred to a fresh Eppendorf tube and dried down to 100 µl. The resulting oligonucleotide was concentrated by ethanol precipitation, and re-suspended in TE buffer (Section 2.7.6).

Table 2.10: Buffers for denaturing PAGE

Buffer	Composition
Electrophoresis running Buffer (1 x TBE buffer)	89 mM Tris- HCl, 89 mM boric acid, 0.2 mM EDTA pH 8.3
Formamide loading Buffer	80% deionised formamide, 10 mM EDTA pH 8.0, 1 mg/ml xylene cyanol, 1 mg/ml bromophenol blue

2.14.3 Native PAGE for fluorescent electrophoretic mobility shift assay (fEMSA)

The binding activities of TALERs were analysed on non-denaturing PAGE. This allowed the detection of TALER-DNA non-covalent bound complexes. These gels were usually of 5% polyacrylamide made from a 30% acrylamide solution (37.5: 1 ratio of acrylamide to bisacrylamide). Sometimes, the gels were supplemented with 5 - 10% glycerol to stabilize the complexes. Gels were pre-run for 30 minutes at 200 V constant voltage before loading the samples, and for 3 - 5 h at 200 V constant voltage after the samples were loaded. All electrophoretic separations were at 4 °C. The electrophoresis running buffer used here was the 1 x TB buffer.

Table 2.10: Buffers for native PAGE

Buffer	Composition
Electrophoresis running Buffer (1 x TB buffer)	89 mM Tris- HCl, 89 mM boric acid

2.15 *In vitro* binding reactions

TALER binding reaction procedure was modified from Arnold *et al.* (1999). Appropriate pairs of PAGE-purified oligonucleotides (Section 2.14.2) were annealed as described in Section 2.7.1, with an excess of the bottom-strand oligonucleotides (1.2x). The top-strand oligonucleotides were designed and ordered from Eurofins Genomics as DNA labelled on the 5' end with the Cy5 fluorescent dye. For the binding assay, 50 nM of Cy5-labelled double-stranded DNA substrates were dissolved in a buffer that contained 20 mM Tris-HCl (pH 7.5), 10 µg/ml poly(dI/dC), 0.002% bromophenol blue and 1% Ficoll. The standard binding reaction samples were made of 20 µl of this binding buffer to which 2.2 µl of TALER protein at the appropriate concentration was added. The samples were mixed and loaded onto a 5% native polyacrylamide gel as described in Section

2.14.7. The gels were run at 200 V (constant voltage) for 3 - 5 hours. After electrophoresis, the gels were scanned using a Typhoon™ FLA 9500 biomolecular imager (GE Healthcare Life Sciences) using 633/670 nm for Cy5 filter.

2.16 *In vitro* recombination reactions

To analyse *in vitro* recombination activity of CRs, the proteins were diluted with 1× chimaeric recombinase dilution buffer (20 mM Tris-HCl pH7.5, 1 mM DTT, 1 M NaCl, 50% glycerol) to 2 µM or the required concentration and aliquots were stored indefinitely at -20 °C. There were two types of *in vitro* reactions analysed in this work: standard recombination reactions and cleavage reactions.

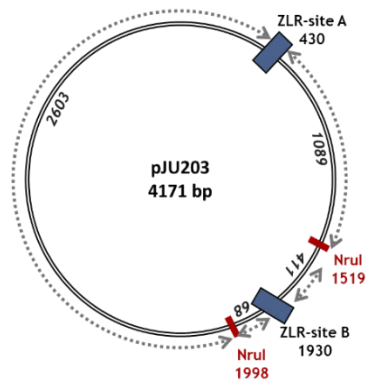
Standard recombination reactions were performed in a buffer containing 50 mM Tris-HCl pH 8.2, 10 mM MgCl₂ and 0.1 mM EDTA. A typical reaction (55.5 µl) contained 1.25 µg of plasmid DNA and 5.5 µl of diluted CR. Reactions were carried out at 37 °C for 1 to 2 hours. The reactions were terminated by heating at 80 °C for 5 minutes. To analyse the recombination products by restriction digest, a 20-µl aliquot was digested with an appropriate restriction enzyme (AlwNI or NruI) for 45-90 minutes at 37 °C. Digested and undigested samples were treated with 0.25 volume of SDS/K loading buffer (SDS loading buffer with protease K added at 1 mg/ml) and incubated at 37 °C for a further 30 minutes before analysis by agarose gel electrophoresis. The expected product sizes from recombination are indicated in Figure 2.10 and Figure 2.11.

Cleavage reaction conditions of 15-40% ethylene glycol (EG) allow the analysis of site-specific CR cleavage activity through the accumulation of cleaved intermediates (Johnson & Bruist, 1989; Boocock *et al.*, 1995). Cleavage reaction buffers were similar to the standard recombination buffers with the inclusion of 40% EG and the exclusion of MgCl₂. Cleavage reactions were typically carried out for 30 minutes and terminated by the addition of 0.25 volume of SDS/K loading buffer with incubation at 37 °C for a further 30 minutes. Reaction products were analysed by agarose gel electrophoresis. No restriction digests were carried out here.

2.17 Computer/Software

Molecular graphics were generated from PDB coordinates using the Pymol software (www.pymol.org). Plasmid sequences and cloning strategies were designed and annotated using the APE (A Plasmid Editor) and SnapGene viewer softwares. Sequence alignments were carried out using the ClustalOmega tool (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) while wild-type recombinase sequence alignments were analysed using Jalview. Structural predictions of CR mutants were designed using the Phyre2 and I-TASSER platforms. Resolvase, Zinc finger, and TALE protein sequences and data were obtained from the PDB and UNIPROT websites. Zinc finger protein design for the HIV CR_TATA_target site was carried out on the ZIFIT software (Sander *et al.*, 2010). The HIV-LANL database was the source of all HIV sequences analysed (HIV.lanl.gov). Multiple tools on the HIV-LANL database were used for HIV data analysis.

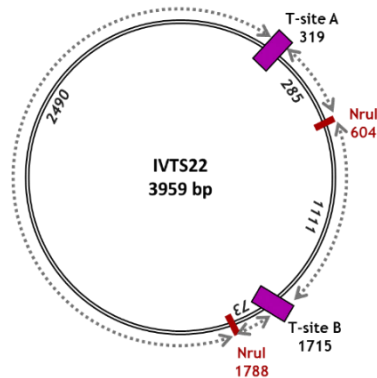
A *In vitro* recombination substrate plasmids



UR - Unrecombined:	479, 3692
RP - Resolution product:	1500, 2671
IP - Inversion product:	1157, 3014

CP- Cleavage products	
Double-site cleavage:	68, 411, 1089, 2603
Single-site cleavage at T-site A:	479, 1089, 2603
Single-site cleavage at T-site B:	68, 411, 3692

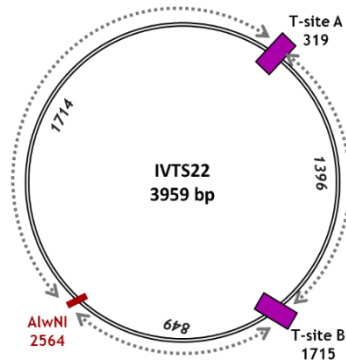
B



UR - Unrecombined:	1184, 2775
RP - Resolution product:	1396, 2563
IP - Inversion product:	358, 3601

CP- Cleavage products	
Double-site cleavage:	73, 285, 1111, 2490
Single-site cleavage at T-site A:	285, 1184, 2490
Single-site cleavage at T-site B:	73, 1111, 2775

C

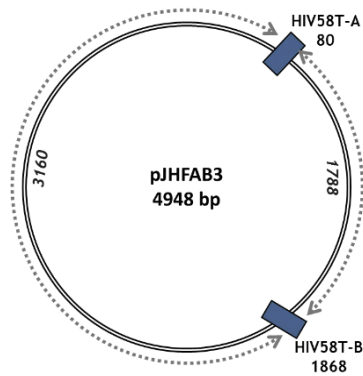


UR - Unrecombined:	3959
RP - Resolution product:	1396, 2563
IP - Inversion product:	3959

CP- Cleavage products:	
Double-site cleavage:	849, 1396, 1714
Single-site cleavage at T-site A:	1714, 2245
Single-site cleavage at T-site B:	849, 3110

Figure 2.10: Expected fragment sizes from *in vitro* recombination activity. NruI or AlwNI restriction digests were usually used to analyse recombination products. The possible products of complete intramolecular recombination are resolution and inversion products. Some of the substrate plasmids might also be left unrecombined while incomplete recombination activity will lead to single or double-site cleavage of the substrate plasmids. As indicated in Figure 2.6, the cloning of *in vitro* recombination substrate plasmids results in two possible plasmid architectures based on the orientation of the kanamycin resistance gene cassette. The expected product sizes from an NruI digest of recombination products from each architecture is shown in A. and B. Kanamycin resistance gene cassette orientation does not affect AlwNI digest analysis. Expected product sizes are indicated in C. Do note that these are only indicative sizes and specific sizes of substrate plasmids and recombination products vary slightly.

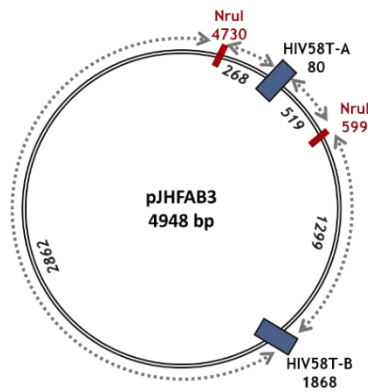
A *In vivo* recombination substrate plasmids



UR - Unrecombined:	4948
RP - Resolution product:	1788, 3160
IP - Inversion product:	4948

CP- Cleavage products	
Double-site cleavage:	1788, 3160
Single-site cleavage at T-site A:	4948
Single-site cleavage at T-site B:	4948

B



UR - Unrecombined:	787, 4161
RP - Resolution product:	1788, 3160
IP - Inversion product:	1567, 3381

CP- Cleavage products	
Double-site cleavage:	268, 519, 1299, 2862
Single-site cleavage at T-site A:	268, 519, 4161
Single-site cleavage at T-site B:	787, 1299, 2862

Figure 2.11: Expected fragment sizes from *in vitro* recombination activity on *in vivo* recombination substrate plasmid. Where recombination products from *in vitro* recombination activity was analysed using MacConkey agar assay (Section 5.2.4), the *in vivo* recombination substrate plasmid was used as a substrate. The expected fragment sizes without restriction digest (A) and with a NruI digest (B) are indicated. Diagnostic AlwNI restriction digest was not appropriate with pJHFAB3 as the full HIV T-sites contain AlwNI restriction sites.

Chapter 3: Engineering the N-terminal catalytic domain

3.1 Research Strategy: Modular Engineering

To generate active CRs that recognize and catalyse the excision of the HIV-1 proviral DNA, it is necessary to alter the recognition specificity of both the DNA-binding domain and the recombinase catalytic domain. Considering the modular structure of serine recombinases and previous success in the broadening of the catalytic activity of ZFRs to non-cognate sequences, the DNA-binding domain and the catalytic recombinase domain can be unlinked and engineered separately. A functional ‘designer’ CR can then be reassembled from these parts (Fig. 3.1). This modular engineering approach is a significant aspect of synthetic biology that thrives on the functional independence of the domains of some proteins, allowing ease of manipulation for the design of novel functionalities (Maervoet and Briers, 2017).

Here, the strategies taken to generate active HIV-targeting recombinase catalytic domain variants and the characterization of these mutants are exhaustively covered. Chapter 4 covers the optimization and design of the DNA-binding domain. The HIV-targeting catalytic domain variants are then fused with the optimal DNA-binding domain to generate functional HIV-targeting CRs as described in Chapter 5.

Since the optimal architecture of Tn3 zinc finger recombinases has been defined (Prorocic *et al.*, 2011; Akopian *et al.*, 2003), the HIV catalytic domain will be designed in this context. A fixed DNA-binding protein, Zif268 will be used to modulate DNA binding activity guiding the ZFR to the modified target site as previously described by Prorocic *et al.*, 2011. The designed catalytic domain can be easily moved to any other CR DBD system such as TALER or CRISPR-recombinase

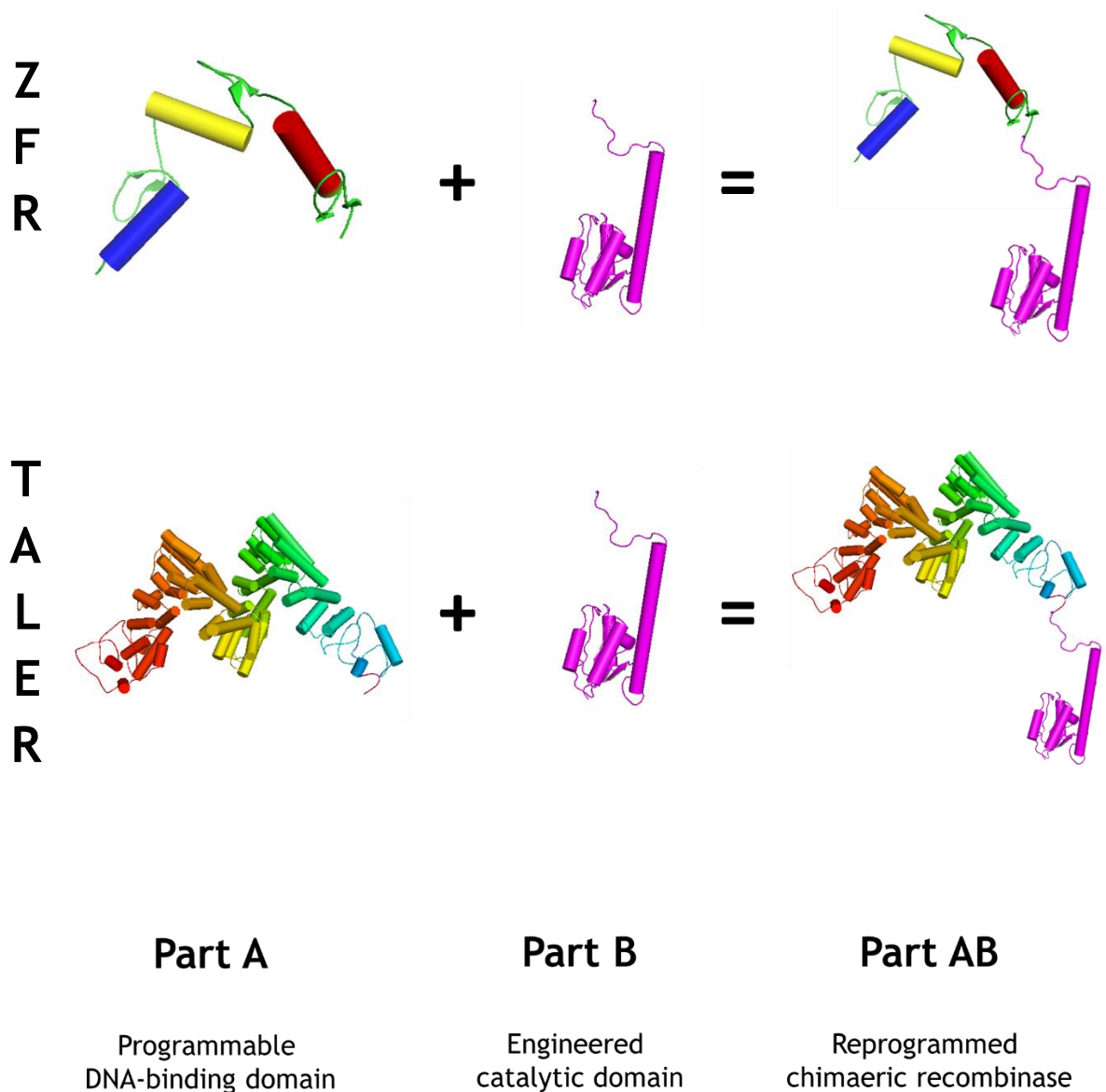


Figure 3.1: Modular Engineering of Chimaeric recombinases. Small serine site-specific recombinases can be fully reprogrammed to modulate recombination activity on non-cognate sites by uncoupling their DNA-binding domain from the catalytic domain (Part B). The catalytic domain can be engineered using random and rational mutagenesis approaches. The engineered catalytic domain can then be fused to a programmable DNA binding domain (Part A). The programmable DBDs usually consist of repeat modules, fingers for ZF proteins and RVDs for TALEs. The new protein (Part AB) will then carry novel target site specificity and should retain the properties of its two parts.

3.2 Introduction: Engineering the Catalytic domain

Previous work in redirecting the site-specificity of recombinase catalytic domains has either focused on switching sequence specificity between two recombinases like Gin invertase and Tn3 resolvase (Gaj *et al.*, 2010) or on targeting sequences that are similar to the recognition sequence of the starting enzyme (Proudfoot *et al.*, 2011; Gersbach *et al.*, 2010; Gordley *et al.*, 2007). In the former strategy, sequence and structural information is available and so rational design approaches can be applied by shuffling mutations between the two proteins or generating chimaeras. The latter strategy does not pose huge selective challenges either, since a large evolutionary step request is not posed to the enzyme. However, in the therapeutic application of CRs for genomic excision on novel sites, such flexibility in target site selection might not be possible. In this research work, this stringency applies as we are confined by restrictions defined by the HIV-1 proviral DNA and the human genome.

Despite the amount of work done in attempting to identify the determinants of serine recombinase catalytic domain recognition specificity using approaches such as saturation mutagenesis, structure-guided reprogramming and stringent substrate-linked directed evolution techniques, not much consensus has been reached on the residues implicated in direct and indirect DNA readout (Sirk *et al.*, 2014; Gersbach *et al.*, 2010; Gordley *et al.*, 2007; Burke *et al.*, 2004). As the serine recombinases are fairly flexible proteins, not all the residues that make contact with DNA during the recombination process can be identified in available crystal structures. In fact, two crystal structures of $\gamma\delta$ resolvase available- 1GDT and 1ZR4- showing the presynaptic dimer complex and the post-cleavage synaptic tetrameric complex respectively, have markedly different tertiary and quaternary structures with several residues in different positions and distances from the DNA (Li *et al.*, 2005) (Fig. 3.2). Due to this structural flexibility, complete rationality in the reprogramming of recombinases based on structural information might be unachievable for now. A combination of random and rational engineering approaches has been applied in this work to design active Tn3 NM resolvase-based catalytic domain variants that target the HIV-1 proviral DNA for mock genomic excision in bacterial cells.

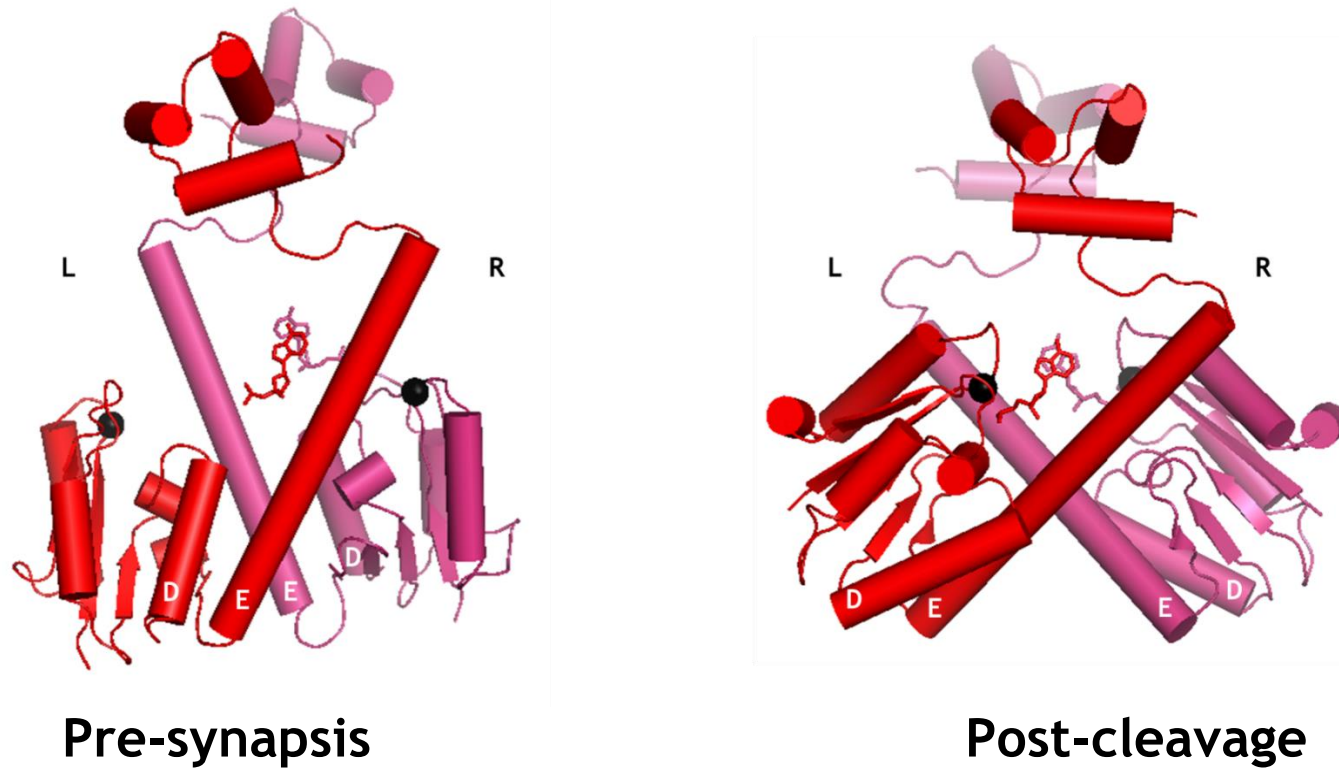


Figure 3.2: Flexibility of $\gamma\delta$ (Tn3) resolvase. Resolvase monomers are shown as red and pink cartoons, with the Ade20 (which carries the scissile phosphate) shown as sticks and coloured according to the corresponding monomer. S10 residues are shown as black spheres. Significant conformational changes can be observed as the protomers move from the pre-synaptic dimeric form (PDB:1GDT- Yang and Steiz, 1995) to the post-cleavage form (PDB: 1ZR2- Li *et al.*, 2005) before strand exchange. According to Li *et al.*, (2005), The two E-helices move about 55° apart and the D-helices also move relative to their own E-helix. The globular region of each monomer moves closer to the scissile phosphate. (Adapted from Li *et al.*, 2005)

3.3 Results

3.3.1 Target site selection

Unlike in nuclease-based genome target site design where a myriad of web-based tools is available for target site selection, CR-based genome engineering is in its infancy and no such tool is currently available. In defining a suitable HIV-1 target site for CRs, these factors were considered as critical hallmarks:

1. The sequence should *ideally be present in direct repeat*, flanking the whole length of the provirus. This is so that limited number of CRs will need to be designed, reducing the potential for off-target activities. The HIV-1 long terminal repeats (LTRs) meet this target site requirement.
2. The sequence must be of extreme importance and *critical to survival of the provirus* such that the target site can elude viral mutagenesis and resistance until all host cells are cleared.
3. The sequence must be *clinically relevant* and present in multiple HIV-1 subtypes in the ‘major’ group, M (especially subtypes A, B and C that constitute over 70% of global HIV-1 incidents globally (Hamelaar, 2012)).
4. The sequence should **not be present on the human genome**.

After reverse transcription of the HIV-1 viral ssRNA, the resulting proviral DNA is flanked by 640-bp sequences called long terminal repeats (LTRs) (Fig. 3.3). This LTR arrangement provides a naturally-occurring direct repeat, an important sequence arrangement for excision by serine recombinases (Section 1.4). These LTRs serve as focal points for viral gene regulation. They are also involved in the integration of HIV into the host genome and contain several binding sites for host and viral factors implicating them in disease virulence and pathogenesis (Krebs *et al.*, 2002). The sequences within the 5' LTR also serve as the core HIV-1 promoter.

The TATA box is a part of the HIV-1 core promoter. The promoter also consists of three tandem binding sites for a transcription factor, Sp1 and an enhancer region containing two NF- κ B binding sites and the TAR element (Fig. 3.3). The TATA box motif, located 30 bp from the transcriptional start site, is highly conserved across several HIV-1 clades showing only a single polymorphism (Karn and Stoltzfus,

2012). There are two main forms of the TATA box motif, one with a central CATATAAGC sequence which is present in almost all HIV-1 subtypes and the second with CATAAAAGC (van Opijnen *et al.*, 2004). This second TATA variant is characteristic of the slowly-replicating HIV subtype E and the AG circulating recombinant forms (CRFs) (Section 1.2). Transcription of HIV-1 genes is initiated by the binding of the TATA-binding protein (TBP) to the TATA box starting the assembly of the RNA Polymerase II pre-initiation complex (PIC). This leads to low-level generation of viral transcripts and translation of proteins such as the transcriptional activator, Tat. Tat-mediated interaction with the RNA structural feature of the TAR element as well as host cellular factors drive increased proviral DNA transcription and elongation using a hijacked host RNA Polymerase II.

The high sequence conservation of the TATA box motif and its neighbouring sequences is inferred to be related to its importance in Tat activity as well as to viral replication and dissemination (Wilhelm *et al.*, 2012; Centlivre *et al.*, 2005). Minor nucleotide changes in the sequences flanking the TATA motif have a significant knockdown effect on Tat-mediated gene expression (Jeeninga *et al.*, 2000; Olsen and Rosen, 1992). The TATA box and its flanking sequences are in effect critical to the activity of the HIV-1 provirus; this meets the second CR target site criterion listed above.

The predominant TATA box central motif CATATAAGC was selected (Fig. 3.3). This motif bears a fortuitous similarity (5 bp of identity: TATAA) to the central sequence of Tn3 *res* site I, where resolvase cuts DNA. A 60-bp LTR sequence was selected, from the closest Sp-1 binding site to the TATA box into the TAR element, with the TATA motif right in the centre. 3492 unique sequences spanning this region were retrieved from the HIV-1 Los Alamos National Library (LANL) database (www.hiv.lanl.gov) mapping from positions 399 to 458 relative to the HXB2 (last date of access - November 2017). An alignment of these sequences showed high conservation across all HIV subtypes in the LANL database. Curated alignments from the LANL databases were also used to generate consensus sequences using the AnalyzeAlign tool (www.hiv.lanl.gov), one containing all 1216 HIV-1 sequences in the curated alignment and the other of 876 sequences representing HIV-1 Major

group, M. These curated alignments contain only one sequence per patient with problematic sequences and biases removed. The consensus logos for these alignments were exactly similar to that shown in Fig. 3.3 for all HIV sequences (data not shown). It is important to note that there is a slight skew in the data on the LANL database (although nearly 50% of global HIV incidences are attributed to subtype C, subtype B with only 12% dominates the LANL DB).

The consensus sequence of the HIV-1 major group alignment was selected as the TATA target sequence (CR_TATA_target) for the experiments to be presented here. A BLAST search of this sequence against the LANL database and the NCBI database showed 100% match to multiple HIV-1 isolates across several countries, years and subtypes. CR_TATA_target thus meets the third critical hallmark of the target site requirement listed above. A BLAST search of the selected 60-bp TATA target sequence against the human genome on the NCBI database yielded no direct or high sequence similarity with any region of available sequences of the human genome. Reducing the search stringency significantly to identify somewhat similar regions only yielded 6 hits of over 176,974 human genomic sequence reads in the database and of 3 billion nucleotides of the human genome. Accounting for duplication of the sequences in the database, validated hits were 3. The maximum similarity of the top hit observed was to 25 nucleotides spaced by 4 gaps on Human chromosome 20 with an E-value of 1.2. This similarity spanned across the catalytic domain target site and since sequence specificity will be directed by the DNA-binding domain, there is little concern for such off-target activity. CR_TATA_target meets the final requirement for excising HIV-1 provirus from the human genome. As the central TATA motif possesses some sequence similarity to the crossover sites of some natural serine recombinases such as Tn3 resolvase and Sin resolvase, the task of designing active catalytic domains to target the site seemed achievable with readily available resources in our research group.

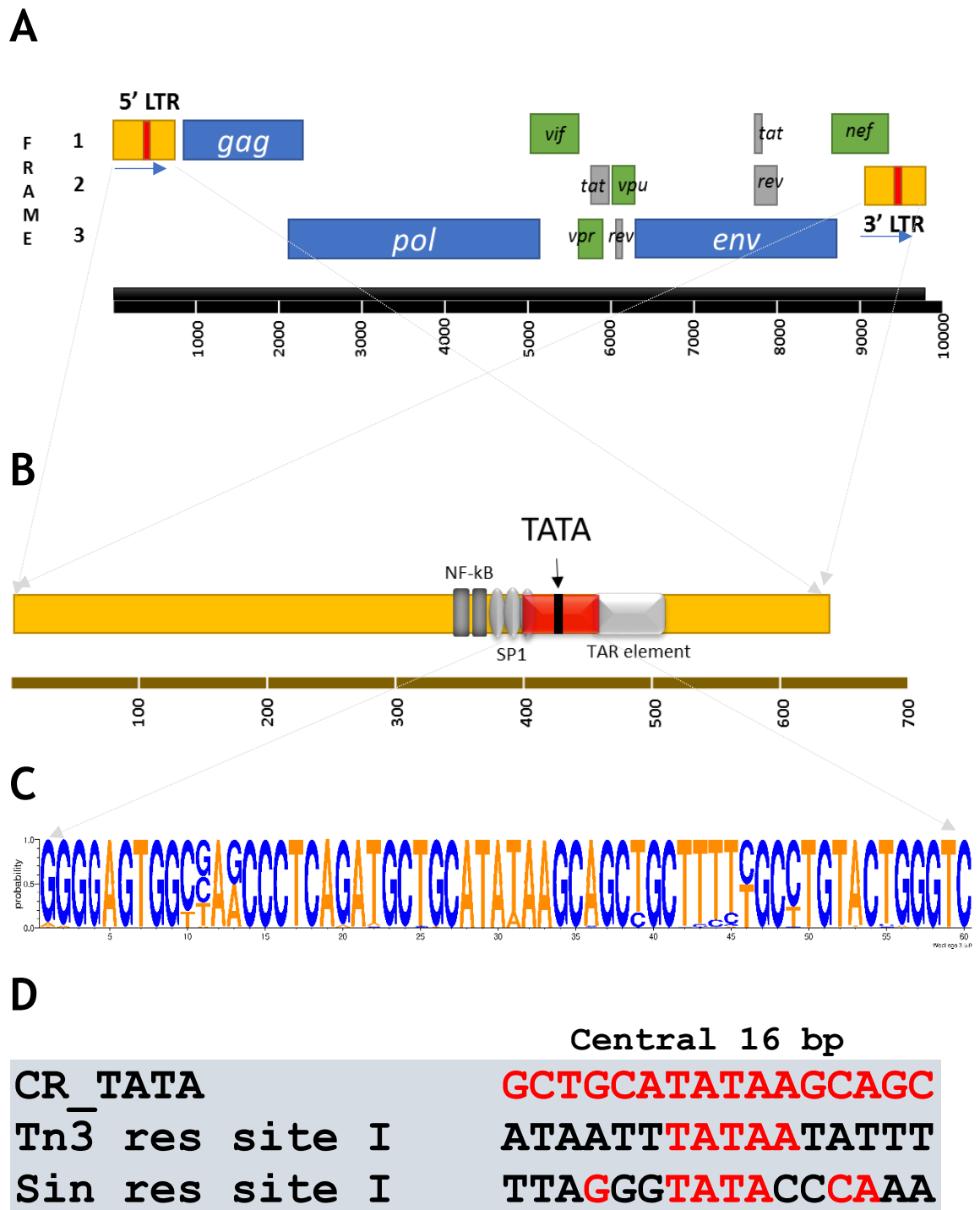


Figure 3.3: The gene map of HIV-1 and the selected CR_TATA_target. A. The LTR is present in direct repeat flanking the coding region of the proviral DNA. B. The core promoter consists of 2 Nf-kB binding sites, 3 SP1 binding sites, the TATA box. The TAR element and the region spanned by the CR_TATA_target is shown in red. C. The consensus sequence logo of the 60-bp CR_TATA_target site from 3492 HIV-1 sequences in the LANL database as generated from the AnalyzeAlign tool (www.hiv.lanl.gov). D. Sequence similarity between the central 16-bp of the CR_TATA_target and the natural recombination sites of Tn3 resolvase and Sin resolvase. Nucleotides similar to the CR_TATA_target sequence are coloured red.

3.3.2 TATA-CR selection target substrate design

Since the TATA-targeting CR catalytic domain will be engineered in a Zif268 ZFR context, a modified Zif268-TATA target site using the 22-bp Z-site architecture reported by Prorocic *et al.* (2011) was designed. The selected CR_TATA target site was reduced to its central 16-bp sequence, flanked on both sides by the 9-bp Zif268 binding sequence with a 3-bp sequence (AGC) spacer on either side. The 3-bp spacer sequence was introduced to extend the Z-site to its optimal 22-bp spacer architecture. It had an added advantage of reducing plasmid instability caused by lengthy inverted repeats in *E. coli* (Fig. 3.4). So that ZFR catalytic domains that are active on each half of the CR_TATA Z-site can be selected, the central 16-bp target sequence was split into two halves, and palindromic sites were constructed with each half (Fig. 3.4). Such increased system modularity in combining two different ZFRs, each one working on half of a target Z-site, has previously been reported by Proudfoot *et al.* (2011). So, three Zif268-TATA target sites were constructed- Zif268-TATA Left-Left (HIV-ZLL), Zif268-TATA Right-Right (HIV-ZRR) and the full Zif268-TATA Left-Right (HIV-ZLR) (Fig. 3.4).

These Z-sites were cloned into substrate plasmids containing a low-copy number pSC101 origin of replication (*ori*), a kanamycin resistance gene (KanR) and a galactokinase (*galk*) gene flanked by two identical ZFR target sites (Z-sites) in direct repeat (Section 2.10.2) (See Chapter 2 for more information on plasmids used in this work). Upon recombination by an active ZFR (expressed from a second plasmid with a *ColEI* origin and an ampicillin resistance gene, AmpR) on the Z-sites, the *galk* gene is excised yielding two circular DNA molecules carrying one Z-site each. The DNA molecule with the *galk* gene, having no origin of replication, becomes lost while the second molecule is retained and replicated as the recombination product. Colonies containing recombination-proficient ZFRs appear white and colonies containing recombination-deficient ZFRs appear red on MacConkey agar supplemented with 1% galactose (see Section 2.11) (Fig. 2.7). The substrate plasmids for HIV-ZLL, HIV-ZRR and HIV-ZLR were called pJU001, pJU002 and pJU003 respectively (Section 2.6).

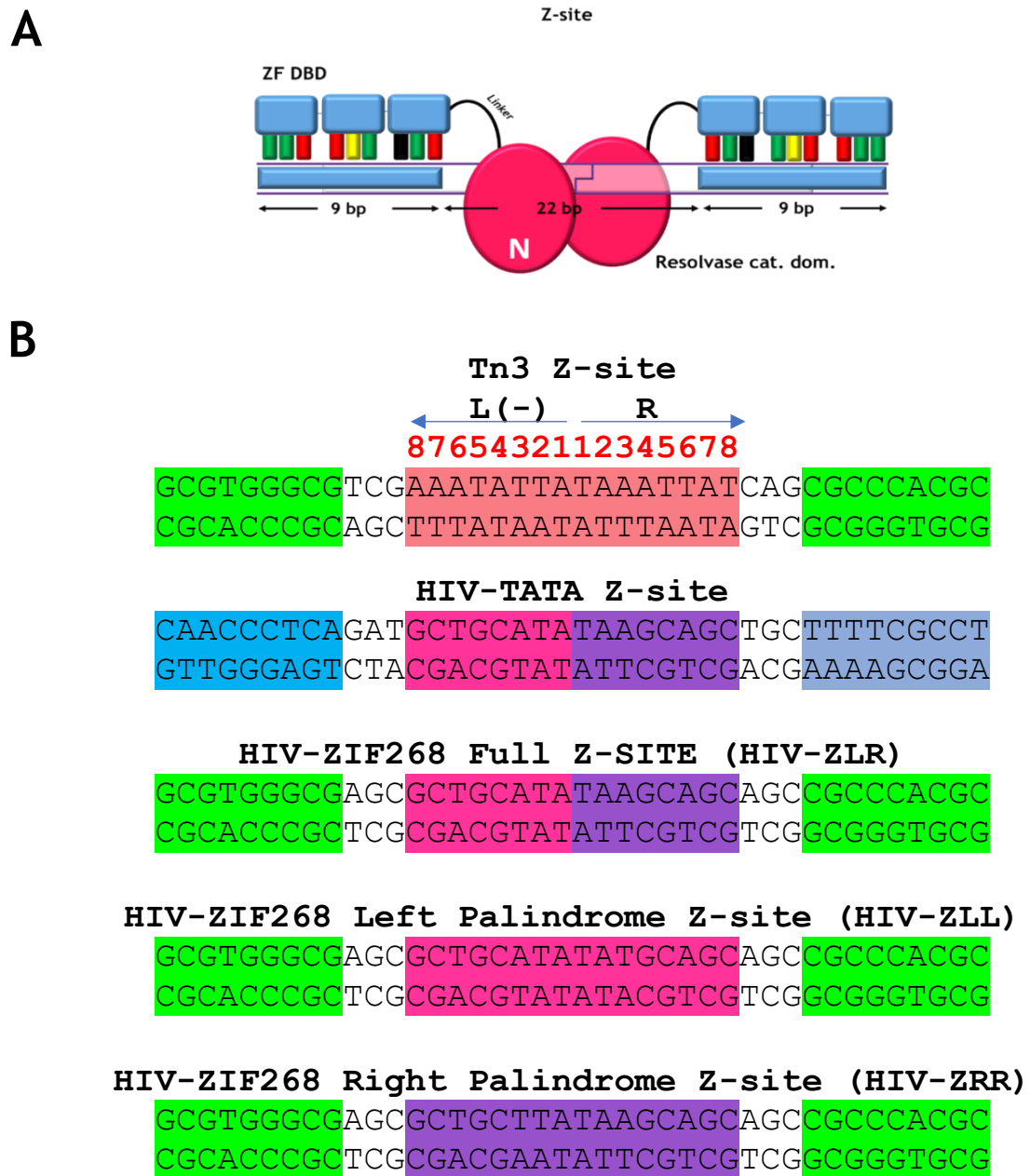


Figure 3.4: The design of the HIV-ZIF268 Z-sites. A. The Z-site architecture as described in Figure 1.17 is shown. B. The sequence of the Z-site from Prorocic *et al.* (2011) with the central 22 bp of Tn3 *res* site I (highlighted in red) flanked by 9-bp Zif268 binding sites (highlighted in green) is shown at the top. The proposed 22-bp spaced Z-site from the CR_TATA_target sequence is shown below it with the left and right zinc finger binding sites highlighted in blue and blue-grey respectively. The central 16-bp sequence of the HIV-TATA Z-site is split into two halves; the right half is highlighted in pink and the left half in purple. To generate the new HIV-Zif268 Z-sites, HIV zinc finger binding sites are replaced with the recognition sequence of a well characterized zinc finger DNA-binding protein, Zif268 to generate the HIV-ZLR site. The sequences of the palindromic HIV-ZIF268 Z-sites, HIV-ZLL and HIV-ZRR are also shown.

3.3.3 Engineering active TATA-ZFRs

Several approaches were taken to generate TATA-ZFRs capable of carrying out recombination on Zif268-TATA sites. These were summarised into four strategies which are explored in the rest of this chapter. Each strategy built on information garnered in the previous one and where possible, available resources in the research group were maximized.

3.3.4 Strategy 1: Analysis of broadened-specificity mutants

Previous work in our laboratory (by Chris Proudfoot) to mutate Tn3 ZFR to recognise and catalyse recombination on the GC-rich Z-Site of Sin resolvase generated many mutants with broadened specificity (Section 1.9.1). As mentioned earlier, the sequence of the CR_TATA target site shows some similarity to the crossover sites of both Tn3 and Sin resolvase, so instead of leaping into intensive library screenings or complicated mutagenic designs, the analysis of the activities of the existing mutants on HIV-ZLL, HIV-ZRR and HIV-ZLR (Section 3.3.2) was considered a good place to start.

Eight Tn3 ZFR mutants were rationally selected from a pool of broadened specificity mutants (Table 3.1) based on previous activity demonstrated on non-cognate sites. The mutants (cMutA, cMutB, cMutC, cMutD, cMutE, cMutF, cMutG and cMutH) were tested for *in vivo* recombination activity on HIV-ZLL, HIV-ZRR and HIV-ZLR substrate plasmids. *E. coli* DS941 cells containing pJU001 (HIV-ZLL), pJU002 (HIV-ZRR) or pJU003 (HIV-ZLR) were chemically transformed with mutant ZFR expression plasmids, selected on MacConkey plates supplemented with 1% galactose and appropriate antibiotics and incubated overnight at 37 °C. The colonies on each plate were scraped off and aliquots of the cell mixtures were grown overnight in L-broth (plus kanamycin to select for substrate/recombination product plasmids) at 37 °C (Section 2.11: approach 1). Plasmid DNA was prepared using a Qiagen miniprep kit and separated using agarose gel electrophoresis (Fig. 3.5). Where present, recombination product bands were excised, and the gel-extracted plasmids were re-transformed into DS941 cells, recovered and sent for sequencing to confirm recombination.

Although there were some white colonies on the MacConkey agar plates from the activities of cMutA, cMutB, cMutC and cMutD on pJU002, the corresponding recombination products were not present when the DNA was prepared and run on an agarose gel (Table 3.2) (Fig. 3.5). Instead, the depletion of substrate plasmids pJU002 and pJU003 was observed on agarose gel; this was interpreted as being due to ZFR-mediated cleavage, based on previous observations by Prorocic (2009). cMutF, cMutG and cMutH showed very minimal activity overall. cMutE yielded no white colonies with pJU001, pJU002 and pJU003; however, there was a recombination band on the gel for all three substrates (quite faint with pJU003). Further sequence analysis revealed that the excised recombination products in most cases were not correct. Only cMutE/RR had recombination products with the predicted sequence in 2 out of 4 samples. The results indicated that these enzymes might be carrying out illegitimate recombination on the plasmid backbone. Subsequent analysis of cMutE on the three substrates revealed that where recombination products were present, substrate depletion was not, and *vice versa* (data not shown). This behaviour was not consistent with typical recombination activity of Tn3 resolvase-based ZFRs.

























Despite this contradiction, it was clear that some of these mutants could carry out recombination on the HIV/Zif268 target sites and although we would not continue with cMutE because of its inconsistent activity, the library that yielded these mutants could be assayed to identify better mutants. Throughout these assays, no single white colony or evidence of substrate depletion was observed with pJU001, suggesting that the HIV-ZLL site was especially refractory to recombination. To optimize the design process, a decision was made to focus on selection of TATA-ZFRs on HIV-ZRR and HIV-ZLR sites first.

Table 3.1: Selected broadened-specificity Tn3toSin ZFR Mutants

	Designation	Mutations (NM +)	Plasmid name
1.	cMutA	I77L R120Q	pCP591
2.	cMutB	R120Q L135R	pCP599
3.	cMutC	I77L E132A	pCP620
4.	cMutD	A115G R120Q E132A	pCP655
5.	cMutE	I77L A115G R120Q E132A	pCP657
6.	cMutF	R120V R130I L135R I138V K139Y F140L	pCP734
7.	cMutG	R120F R130I E132A L135R K136R I138V K139Y F140N	pCP735
8.	cMutH	R120F R130I E132A K136R I138V K139Y F140L	pCP739

These mutants were designed and selected by C. Proudfoot when mutating Tn3 resolvase to recognize Sin resolvase Z-sites. The additional mutations in the Tn3 NM ZFR mutants are shown. The NM mutations from wild-type Tn3 resolvase are R2A E56K G101S D102Y M103I Q105L.

Table 3.2: Recombination activity of selected mutants

		White Colonies			Recombination product band on gel		
		ZLL	ZRR	ZLR	ZLL	ZRR	ZLR
1.	cMutA				No	No	No
2.	cMutB				No	No	No
3.	cMutC				No	No	No
4.	cMutD				No	No	No
5.	cMutE				Yes	Yes	Faint
6.	cMutF				No	No	No
7.	cMutG				No	Faint	Faint
8.	cMutH				No	Faint	No

This table shows the results obtained from the MacConkey agar assay and a summary of the analysis of the resulting DNA on the gel. See also Figure 3.5. The ratio of the pale to red colour in each circle gives an approximation of the proportion of the 'white' to red colonies on the MacConkey agar after overnight incubation at 37 °C; see below for the key (black arrows indicate some white colonies).

*(distinctly reduced number of colonies).



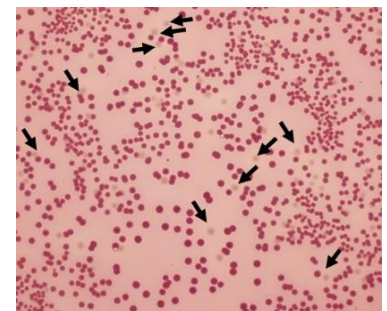
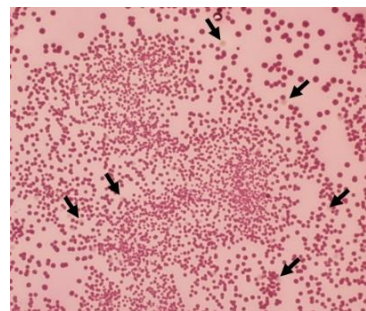
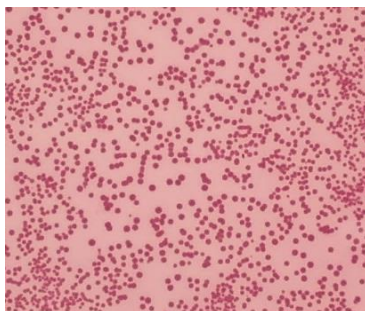
No white colonies



Very few white colonies
(> 10%)



Some white colonies
(10 - 25%)



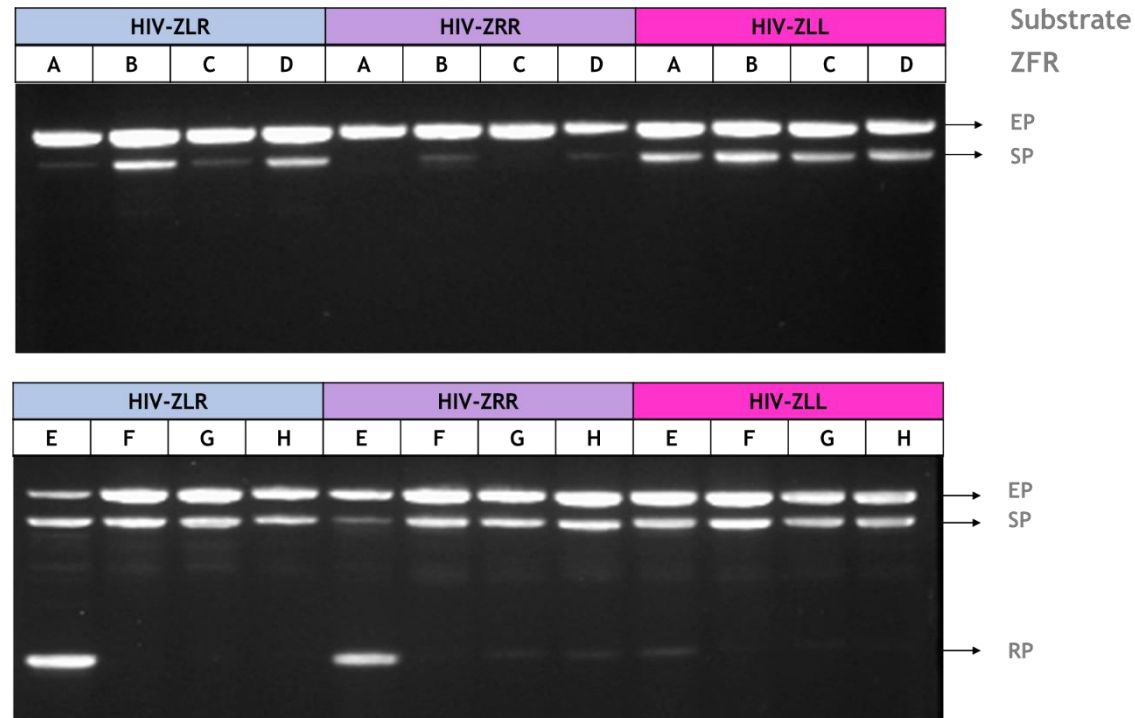


Figure 3.5: Activity of selected broadened-specificity mutants on HIV-ZLR, HIV-ZRR and HIV-ZLL substrate plasmids. The agarose gels here show plasmid DNA prepared from overnight liquid cultures of scraped cells on MacConkey agar plates. The activities of cMutA, cMutB, cMutC and cMutB on pJU001 (HIV-ZLL), pJU002 (HIV-ZRR) and pJU003 (HIV-ZLR) are shown in the top gel and those of cMutE, cMutF, cMutG and cMutH on the three substrates at the bottom. The abbreviations EP, SP and RP refer to Expression plasmid, Substrate plasmid and Resolution product plasmid respectively. Some substrate plasmid depletion can be observed with pJU002 and pJU003 with cMutA, cMutB, cMutC and cMutD; although there is no appearance of resolution products. A band which could be the resolution product is observed with the activity of cMutE on the three substrates. Faint resolution product bands can also be observed with cmutG and cMutH on pJU002.

3.3.5 Strategy 2: Screening of Tn3toSin E-helix libraries

The two rationally designed mutant libraries (created by C. Proudfoot) of ~300,000 Tn3-based Zif268 ZFRs where some of the mutants in strategy 1 were selected from were screened on pJU002 and pJU003 (Section 3.3.4). These libraries also serve as a good pool for selecting ZFRs on the HIV Z-site because combining the specificity for Tn3 and Sin recombination sites should combine the similarities of these sites to the 16-bp central TATA target site (Fig. 3.6). The libraries contain a range of mutations in the E-helix of the Tn3 N-terminal domain with a total possibility of 72,576 variants (Fig. 3.7) although codon bias could lead to proliferation of specific mutations over the others. The libraries were constructed using synthetic oligonucleotides with 2 or 4 alternative nucleotide bases (in equal amounts) at selected positions in the DNA sequence, generating the amino acid variations shown in Figure 3.7C. Both libraries contain the same range of mutations except that the first one has an additional activating mutation (I77L) in the catalytic domain. The libraries are designated 'E-helix +I77L Lib' and 'E-helix-I77L lib'. The I77L mutation was identified as an activating mutation by Burke *et al.* (2004). Proudfoot *et al.* (2011) predicted that the I77L mutation might contribute to recognition of novel sites by relaxing regulatory stringency of catalytic activity.

Variants that yielded white (or 'pinkish') colonies on MacConkey plates after overnight screening of the two mutant libraries were selected and streaked out to obtain single white colonies as described in Proudfoot *et al.* (2011) (Fig. 3.8). Of about 20,000 colonies analysed across this initial library selection, 9 single colonies were selected and streaked out from plates with the transformation of pJU002 with E-helix + I77L Lib, 4 single colonies from pJU002/E-helix - I77L Lib, 10 single colonies from pJU003/E-helix +I77L Lib and 7 single colonies from pJU003/E-helix - I77L Lib. 4 out of the total 30 streaked out to yield white colonies. All four selected mutants were from the E-helix + I77L Lib while the E-helix - I77L Lib yielded no active mutants. DNA was prepared from single colonies to confirm if recombination activity yielded visible resolution product bands on the agarose gel. The recombination products were excised from the gel, purified and confirmed by sequencing as before. The expression plasmids were also sequenced from three samples each, to obtain the mutations that produced the

observed activity. The sequences were aligned with the Tn3 NM resolvase + 177L sequence and the mutations were mapped (Table 3.3). Multiple attempts to select more mutants on the HIV-ZLR substrate using this library were not successful.

```

Sin          -MIIGYARVSSIDQNLERQLDNLKTFGV--EKIFTEKQSGKSVENRPVFQEALNFVRMGD  57
Gammadelta  MRLFGYARVSTSQQSLDIQVRALKDAGVKANRIFTDKASGSSDRKGL-DLLRMKVEEGD  59
Tn3          MRIFGYARVSTSQQSLDIQIRALKDAGVKANRIFTDKASGSSTREGL-DLLRMKVEEGD  59
Tn3NM       MAIFGYARVSTSQQSLDIQIRALKDAGVKANRIFTDKASGSSTREGL-DLLRMKVEEGD  59
           ::*****: :*. *: * : ** ** :***: * **.* :.. : : * . **

Sin          RFVWESIDRLGRNYDEIITETVNYLKEKDVLIMITSLPMMNEVIGNPLLDKFMKDLIIQIL  117
Gammadelta  VILVKKLDRLGRDTADMIQLIKEFDAQGV SIRFI-----DD---GISTDGMGKMVVTIL  111
Tn3          VILVKKLDRLGRDTADMIQLIKEFDAQGVAVRFI-----DD---GISTDGMGQMVVTIL  111
Tn3NM       VILVKKLDRLGRDTADMIQLIKEFDAQGVAVRFI-----DD---GISTDSYIGLMVVTIL  111
           :*.: :*****: :*.: :. :. :.* : : : : : . * : : : : **

Sin          AMVSEQERNESKRROAQQIQVAKEKGVYKGR-----  148
Gammadelta  SAVAQAERQRILERTNEGRQEAMAKGVVFGRRKIDR  148
Tn3          SAVAQAERRRILERTNEGRQEAKLKGIFGRRRTVDR  148
Tn3NM       SAVAQAERRRILERTNEGRQEAKLKGIFGRRRTVDR  148
           : *.: : *.. . * : * * * ** : **

```

Figure 3.6: Sequence alignments of Sin, $\gamma\delta$ and Tn3 catalytic domains. Multiple sequence alignment of the catalytic domain sequence of wild-type Sin, $\gamma\delta$, Tn3 resolvases and Tn3 NM deregulated mutant using the Clustal omega tool (<https://www.ebi.ac.uk/Tools/msa/clustalo/>, 2017). The I77 residue is marked by a red box. The first 148 (142 for Sin resolvase) amino acids are used up to the cut-off point for ZFR design. In the ZFRs used here, the Tn3 resolvase catalytic domain and the Zif268 DBD are connected by a 2-amino acid linker (TS) between residue 148 of resolvase and residue 2 of the Zif268 DBD. Tn3 and Sin resolvases share about 34.5% of sequence conservation in the sequence space analysed.

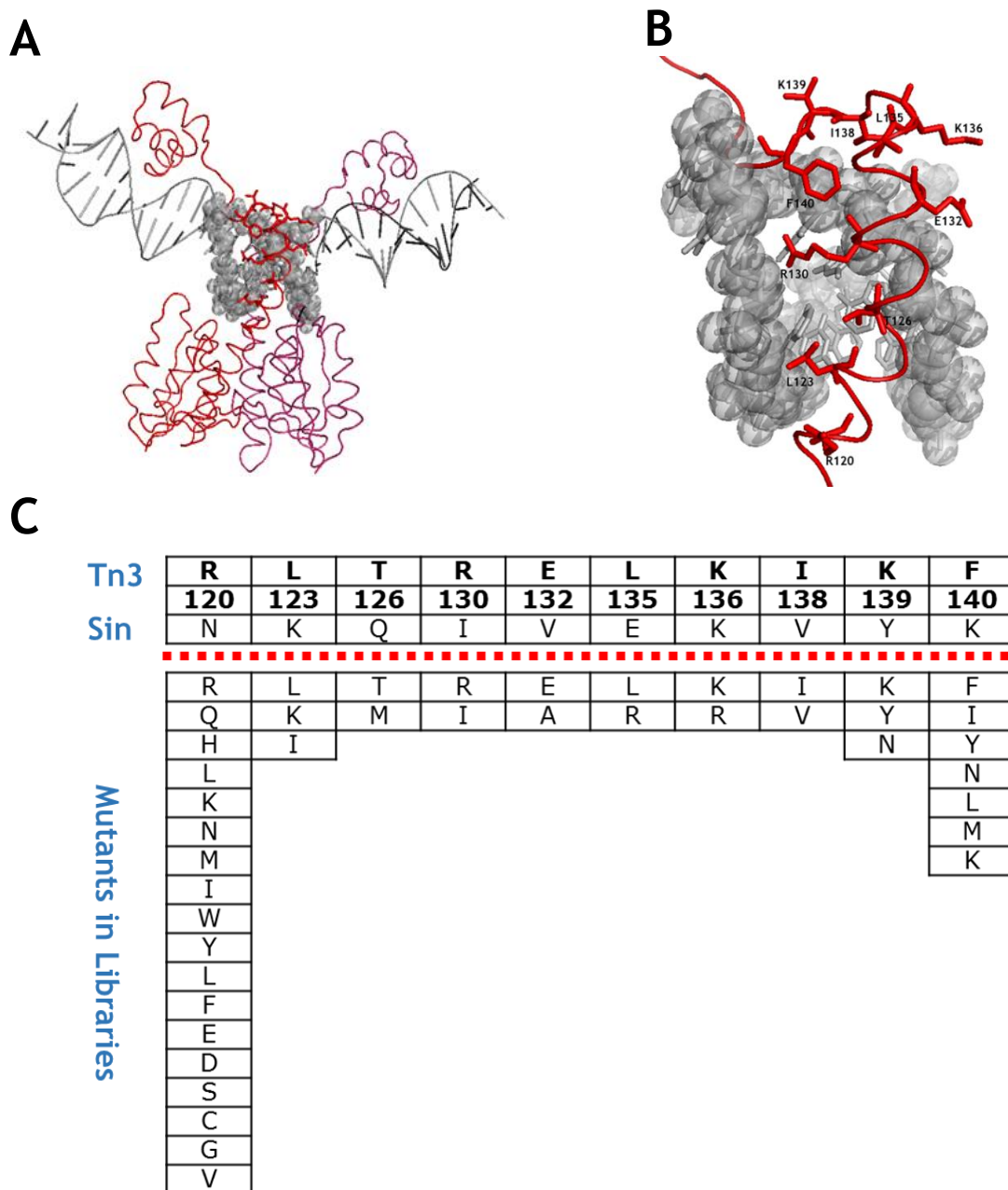


Figure 3.7: E-helix library design. **A.** Cartoon representation of crystal structure of $\gamma\delta$ resolvase bound to site I DNA (PDB: 1GDT - Yang and Steiz, 1995) showing the position of mutated residues in the E-helix libraries made by C. Proudfoot. DNA is shown as grey cartoon and spheres. Resolvase monomers are shown in red and pink. **B.** Closer look at the mutated residues show proximity to DNA. Residues are shown as red sticks. **C.** Mutants in E-helix libraries. The original residue in wild-type Tn3 resolvase is shown in the top row in bold, residue position in the middle and then the corresponding residue in Sin resolvase is shown below. The actual library composition is shown below the red line. Over 72,000 possible combinations of amino acids can be generated from this library. Some of the residues are mutated to that in Sin resolvase and others to previously identified activating mutations.

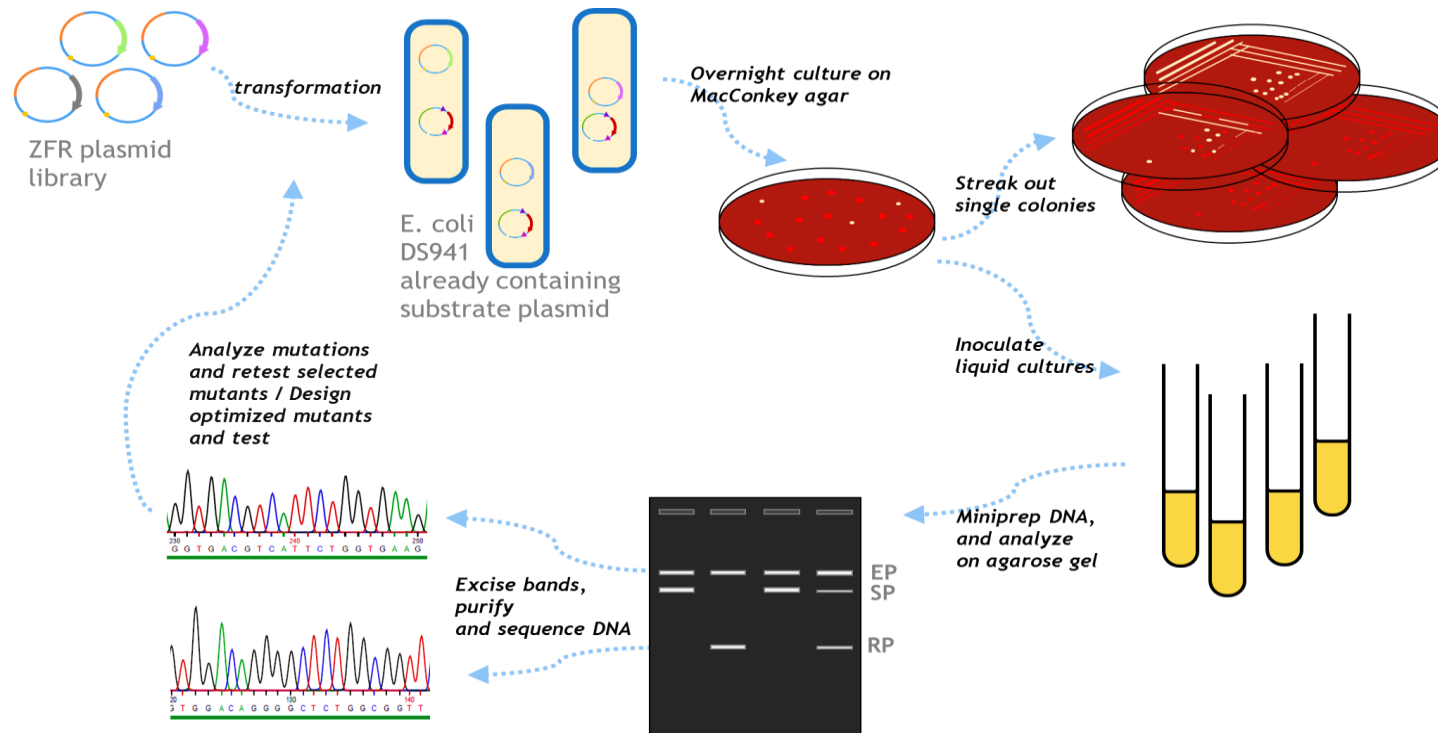


Figure 3.8: Schematic for library screening for active ZFR mutants. The ZFR plasmid library was transformed into competent *E. coli* cells already containing the substrate plasmids. After a 90-minute expression step, the cells are screened overnight on MacConkey agar plates containing 1% galactose and appropriate antibiotics. If present, white colonies are carefully streaked out and simultaneously inoculated into liquid cultures containing a single antibiotic (kanamycin) to select for the resolution or substrate plasmid. DNA is prepared and run on an agarose gel to confirm presence of resolution products. The bands corresponding to resolution products and enzyme plasmids are excised, DNA is purified and sent for sequencing. The sequences are analysed, and the selected mutants are retested on the substrate plasmids to confirm activity.

Table 3.3: 1st generation HIV-Zif268 mutant ZFRs

Designation	MUTATIONS (NM + I77L +)	ISOLATED ON	PLASMID NAME
ZR001	R120D T126M R130I I138V K139N	pJU002	pJUM001
ZR003	R120I T126M L135R I138V F140M	pJU002	pJUM003
ZR004	R120L E132A I138V	pJU002	pJUM004
ZR005	R120F L123I E132A F140N	pJU003	pJUM005

3.3.5.1 *Recombination activity of 1st generation mutants on target sites*

In vivo activity of the selected mutant ZFRs (ZR001, ZR003, ZR004 and ZR005) was analysed on pJU001, pJU002 and pJU003 (See Section 2.11). ZR001 was found to be inactive on any of the substrates. ZR003 and ZR004 showed stable and specific recombination activity on pJU002, but no significant activity was observed on either pJU001 or pJU003. ZR005 was isolated as a dimer gene on its plasmid and when separated into monomers, no significant recombination activity was observed on any of the substrates (data not shown).

In vivo activity of these mutants was also analysed on Tn3 and Sin Z-substrate plasmids, pMP53 and pMP217 respectively (Fig. 3.9, Fig. 3.10) (Section 2.6). None of the mutants was active on the Sin Z-substrate, indicating that the catalytic specificity of the mutants has not been widely broadened. ZR001 was inactive on the Tn3 Z-substrate, suggesting a loss of catalytic function (or DNA recognition). ZR003 and ZR004 showed stable recombination products with the Tn3 Z-substrate, to a greater degree than on RR. However, ZR004 does not completely convert the Tn3 Z-substrate plasmid, implying the beginning of a loss of specificity for the Tn3 site.

The four ZFRs were expressed and purified, and *in vitro* recombination activity analysis was carried out (data not shown). *In vitro* recombination activity was similar to that obtained *in vivo*. ZR004 showed more recombination efficiency than ZR003. However, these proteins did not show as much activity as that of the Tn3 NM ZFR on a Tn3 *in vitro* Z-substrate plasmid (pDWIVS6). This implied that their activity was not optimal.

ZR004 was preferred over ZR003 as it shows a slight loss of function on the Tn3 site and has only three additional mutations (to NM+177L) relative to ZR003's five. The mutations in the E-helix ZR004 were separated and isolated to generate single mutant variants (including the NM + 177L mutations). Mutational analysis of ZR004 using these plasmids indicated that the R120L mutation was essential for recombination on RR and that E132A and I138V improved this activity (data not shown). All three mutations (+NM + 177L) are required for the level of activity ZR004 shows on RR.

No significant recombination activity was obtained on the full HIV-ZLR substrate plasmid (pJU003) and no ZFR mutant showed any activity on HIV-ZLL substrate plasmid (pJU001), either *in vivo* or *in vitro*. The activity of ZR003 and ZR004 on the HIV-ZRR substrate relative to that on Tn3 Z-substrate was also suboptimal. This indicated that although the E-helix mutations contributed to a shift in substrate specificity, there might be a need for non-E-helix activating mutations in the catalytic domain to enable recombination of these new sites. The lack of cleavage or recombination activity on the HIV-ZLL site and very limited activity on HIV-ZLR might indicate a unique feature in the left sequence of the CR_TATA target 16-bp core that deters recombination activity.

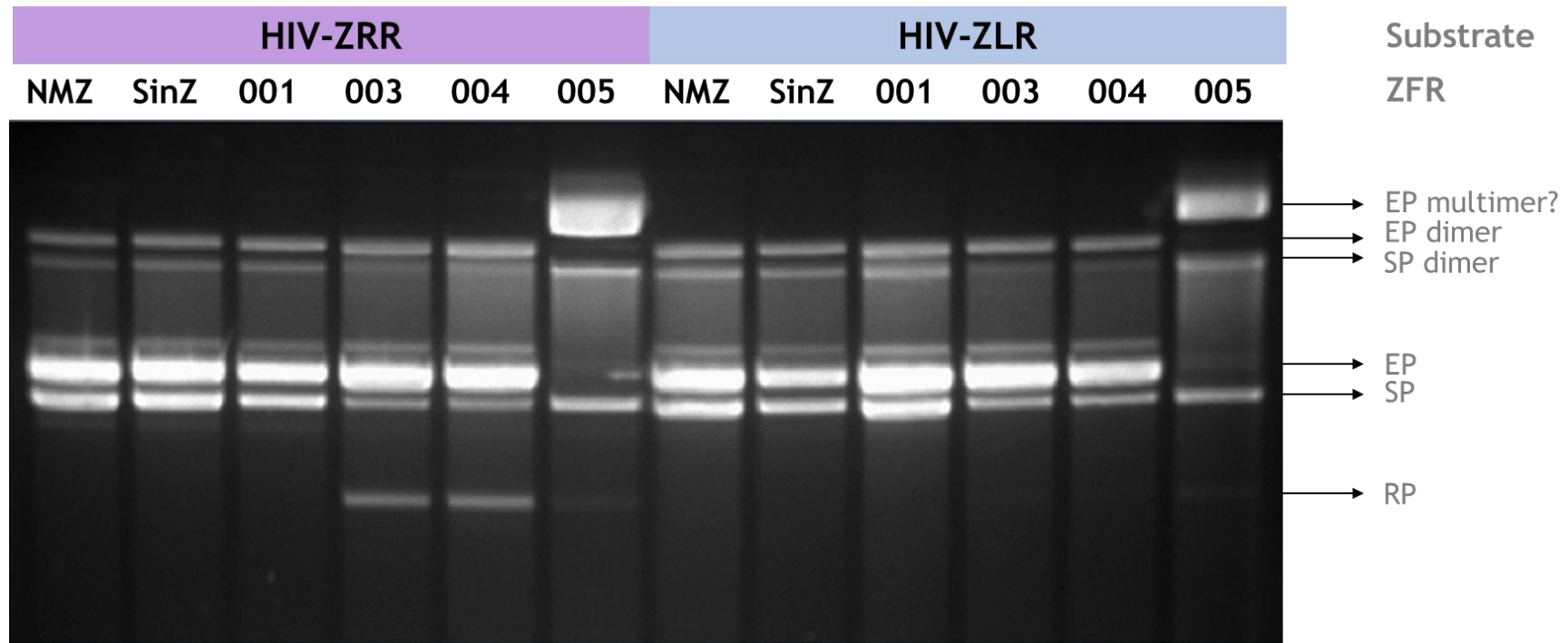


Figure 3.9: Activity of 1st generation ZFR mutants on HIV-ZRR and HIV-ZLR substrate plasmids. Tn3 NM Z-resolvase expression plasmid (pMP59), Sin Z-resolvase plasmid (pMP213), pJUM001, pJUM003, pJUM004 and pJUM005 were transformed into competent *E. coli* cells containing either pJU002 (HIV-ZRR SP) or pJU003 (HIV-ZLL SP). After 90 minutes expression, cells were inoculated into L-broth (containing kanamycin and ampicillin). DNA products recovered from overnight cultures were run on agarose gels. Tn3 NM Z-resolvase and Sin Z-resolvase do not recombine pJU002 or pJU003. ZR001 shows no activity on pJU002 or pJU003. ZR003 and ZR004 yield resolution product bands on pJU002 but not on pJU003. ZR005 turned out to be a plasmid multimer and shows very minimal activity on pJU002 and pJU003.

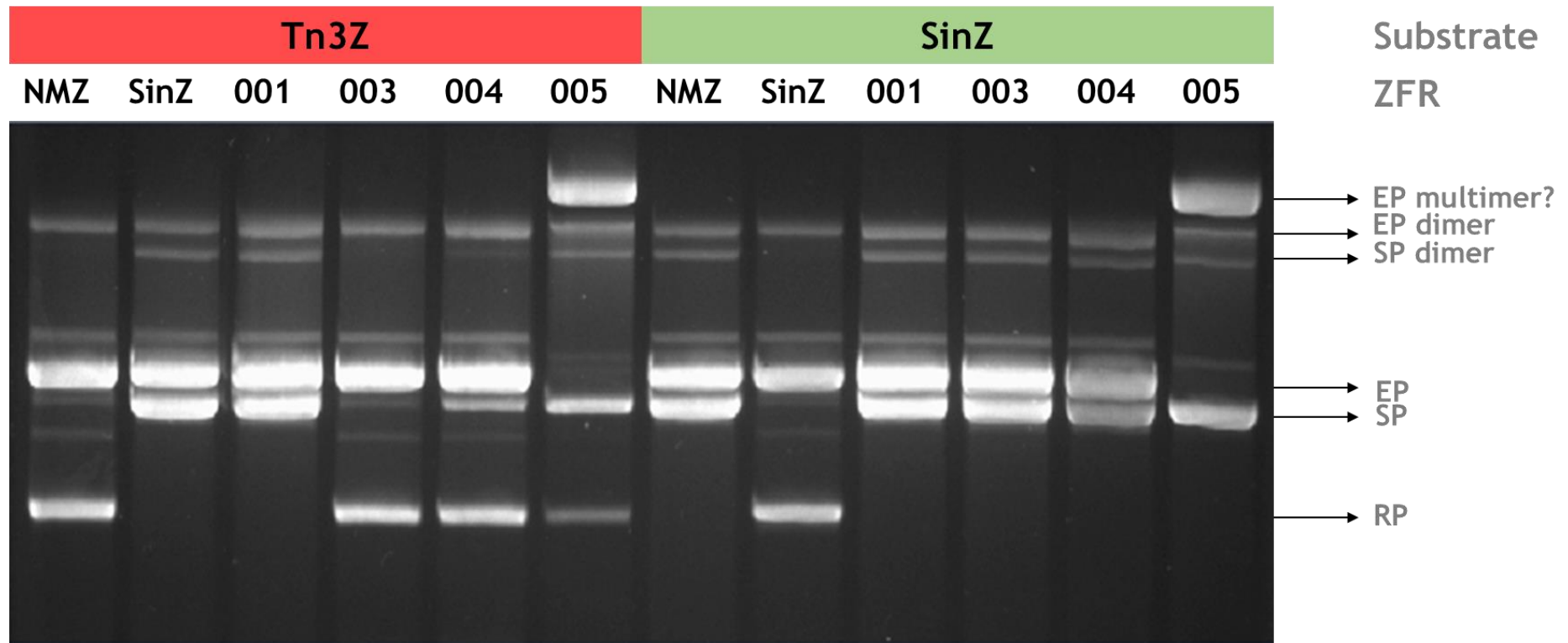


Figure 3.10: Activity of 1st generation ZFR mutants on Tn3 and Sin Z-substrates. Tn3 NM Z-resolvase expression plasmid (pMP59), Sin Z-resolvase plasmid (pMP213), pJUM001, pJUM003, pJUM004 and pJUM005 were transformed into competent *E. coli* cells containing either pMP53 (Tn3 Z-SP) or pMP217 (Sin Z-SP). After 90 minutes expression, cells were inoculated into L-broth (containing kanamycin and ampicillin). DNA products recovered from overnight cultures were run on agarose gels. Tn3 NM and Sin resolvase recombine their own target substrates but not each other's. ZR001 shows no activity on pMP53 or pMP217. ZR003, 004 and 005 yield resolution product bands on pMP53. None of the mutants are active on pMP217.

3.3.6 Strategy 3: Screening of hyperactive mutants on HIV sites

To identify mutations that might improve the catalytic activity of a specificity-shifted mutant ZFR, an available hyper-active ZFR (ZR0122) selected and purified by C. Proudfoot with the mutations K29E G70S V107L (+ NM mutations) was analysed. *In vitro*, this mutant showed very significant cleavage of the HIV-ZRR substrate (data not shown). However, *in vivo* recombination revealed the loss of the substrate plasmid without the appearance of a stable recombination product (Fig. 3.11). This phenotype implied cleavage at the HIV-ZRR site without religation leading to plasmid loss. This substrate depletion was also observed on the HIV-ZLR site substrate. To ‘tame’ the activity of ZR0122, mutational analysis was carried out to narrow down the mutation(s) responsible for its increased cleavage activity (Fig. 3.11). This led to the identification of the K29E mutation as significant for its cleavage activity. A new mutant ZFR (ZR113) with only one non-E-helix mutation, V107F (+NM +I77L) that showed some recombination activity on pJU002 was also identified (Fig 3.11).

Introducing the K29E mutation into ZR0013 resulted in a new mutant, **ZR012** (NM + K29E I77L V107F) that showed almost complete recombination of RR (Fig. 3.12). However, its activity on the HIV-ZLR substrate indicated depletion of substrate without the appearance of recombination product (Fig. 3.13). To check if combining the mutations in ZR004 (Section 3.3.5) and ZR012 would yield HIV ZFR mutants that have an increased activity resulting from shifted specificity and improved catalytic strength, two new ZFR mutants were designed (**ZR014**: NM+ K29E I77L V107F R120L E132A I138V and **ZR015**: NM + K29E I77L R120L E132A I138V). ZR014 has the V107F mutation while ZR015 does not. On pJU002, ZR014 showed marginally better recombination activity than ZR004 and ZR012, with complete substrate depletion and a stronger resolution product band, while ZR015 showed reduced activity compared to ZR012, indicating the importance of the V107F mutation (Fig. 3.14). However, none of these mutants show visible activity on pJU001, while considerable substrate depletion can be observed with pJU003. This conclusively indicates that the left half sequence of the 16-bp core CR_TATA target site poses a greater challenge to recombination than the right half.

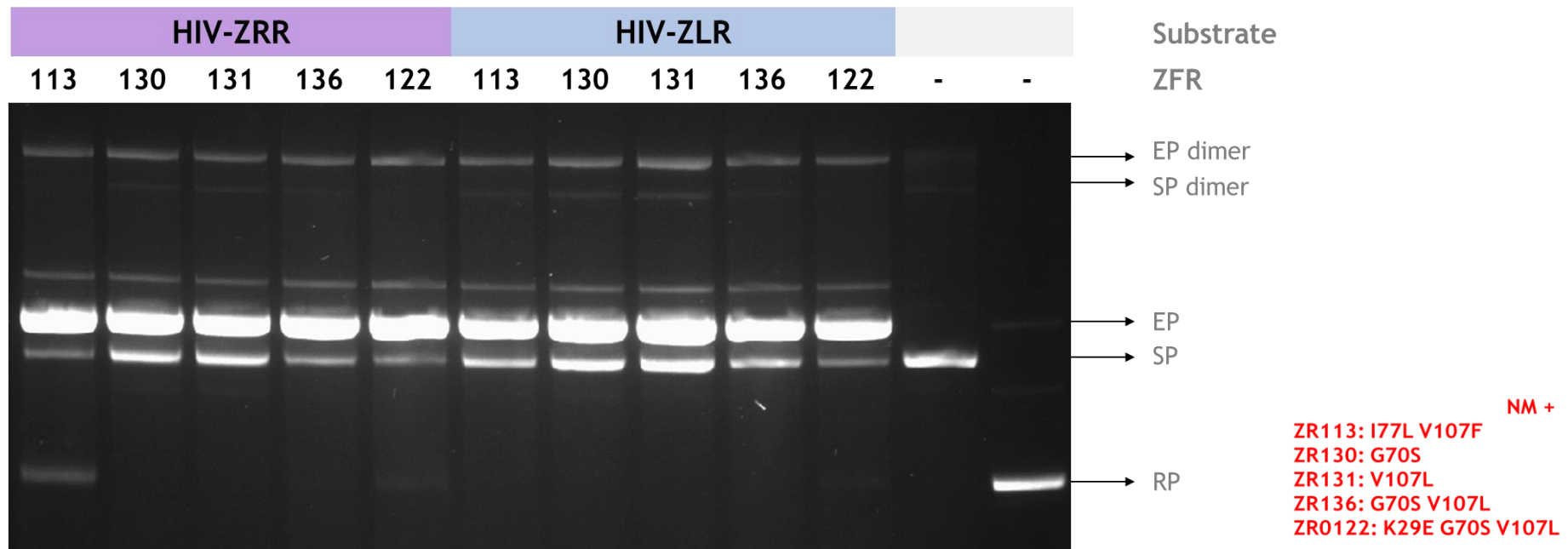


Figure 3.11: Taming ZR0122. The expression plasmids of ZR0122 (pAM0122), its double mutation variant ZR136 (pAM0036), its single mutation variants ZR130 (pAM0030) and ZR131 (pAM0031), as well as that of new mutant variant ZR113 (pAM0013), were used to transform *E. coli* cells containing either pJU002 or pJU003. After 90 minutes expression, cells were inoculated into L-broth (containing kanamycin and ampicillin). DNA products recovered from overnight cultures were run on agarose gels. ZR0122 depletes pJU002 and pJU003 with very minimal appearance of resolution products on pJU002 and none on pJU003. The analysis of single and double mutant variants of ZR0122 should help to identify the mutation(s) responsible for cleavage activity on pJU003. Single mutants ZR130 and ZR131 do not show significant activity. The K29E mutation in ZR0122 seems important for improved cleavage activity. V107F mutant, ZR113 yields resolution products on pJU002 but not on pJU003. pJU002 and a sequenced resolution product plasmid are run as controls to indicate substrate plasmid and expected resolution product sizes on the gel.

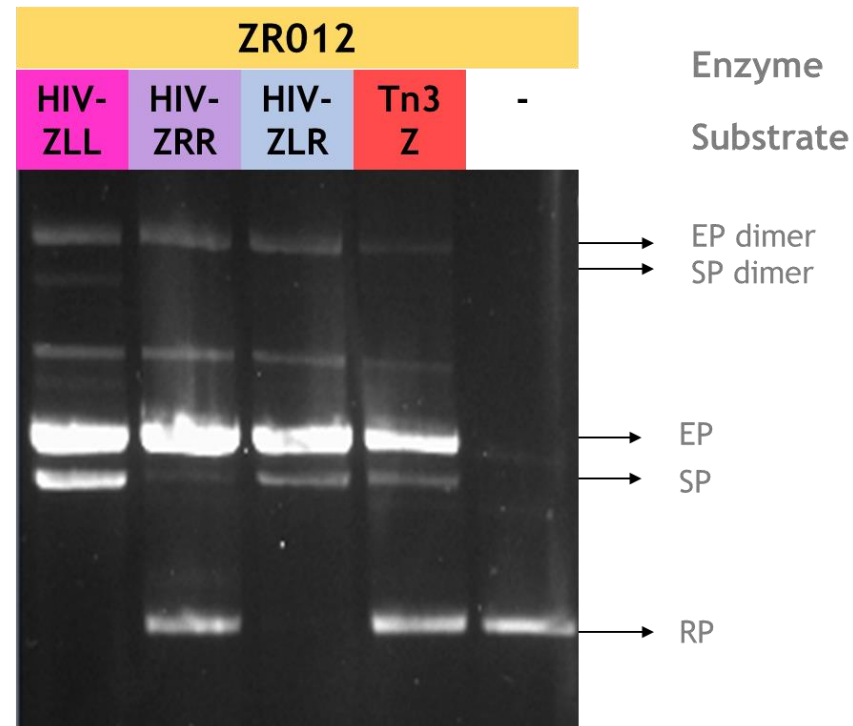


Figure 3.12: ZR012 yields recombination product on HIV-ZRR substrate. The expression plasmid of ZR012 (pJUM012) was transformed into competent *E. coli* cells containing either pJU001, pJU002, pJU003 or pMP53. After 90 minutes expression, cells were inoculated into L-broth (containing kanamycin and ampicillin). DNA products recovered from overnight cultures were run on agarose gels. The activity of ZR012 yields recombination product plasmid on pJU002 and pMP53. There are no recombination products from the pJU001 and pJU003; however, there is a significant reduction in the amount of pJU003. pJU001 does not appear to be depleted. pJU002 and a sequenced resolution product plasmid are run as controls to indicate substrate plasmid and expected resolution product sizes on the gel.

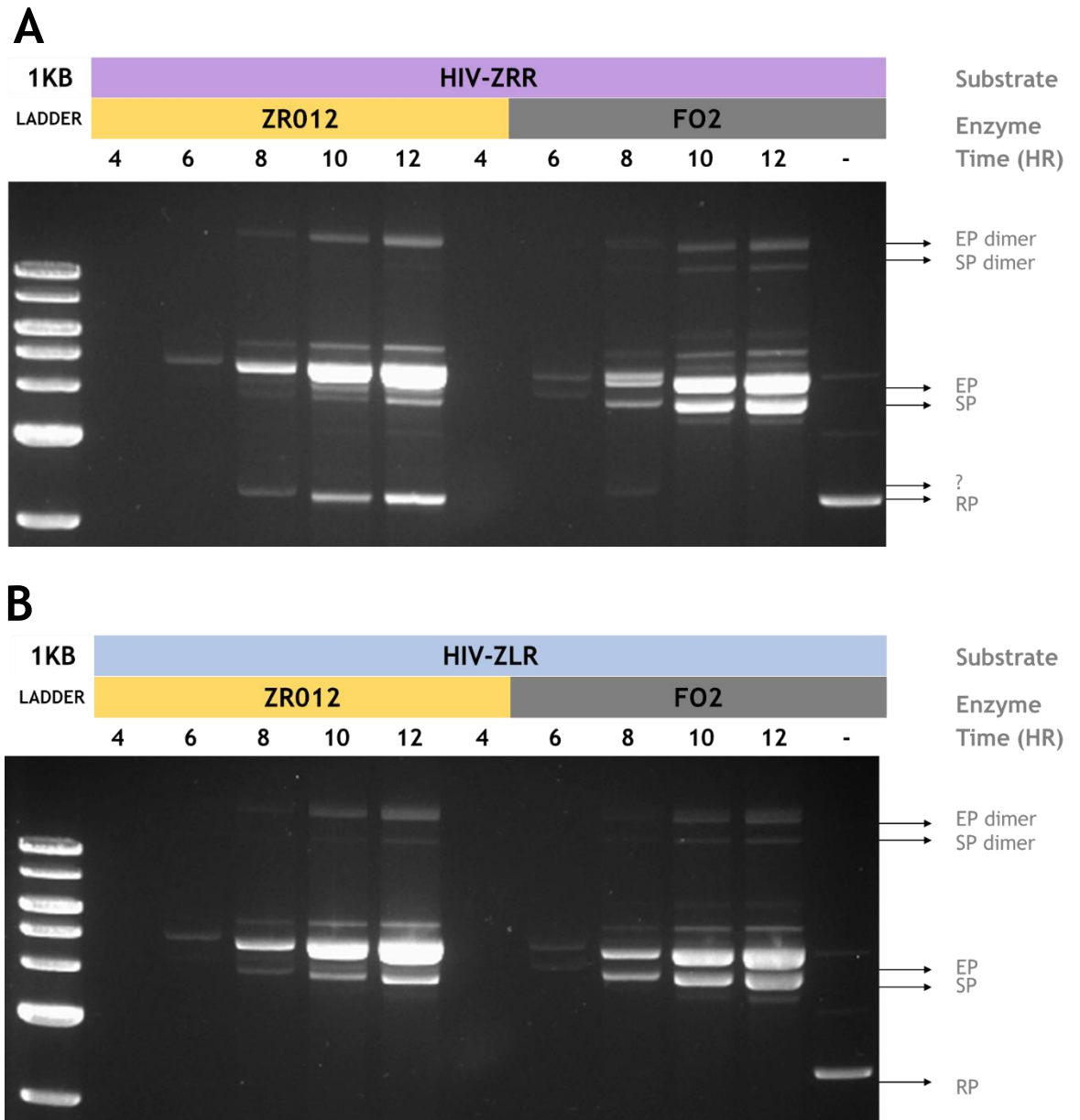
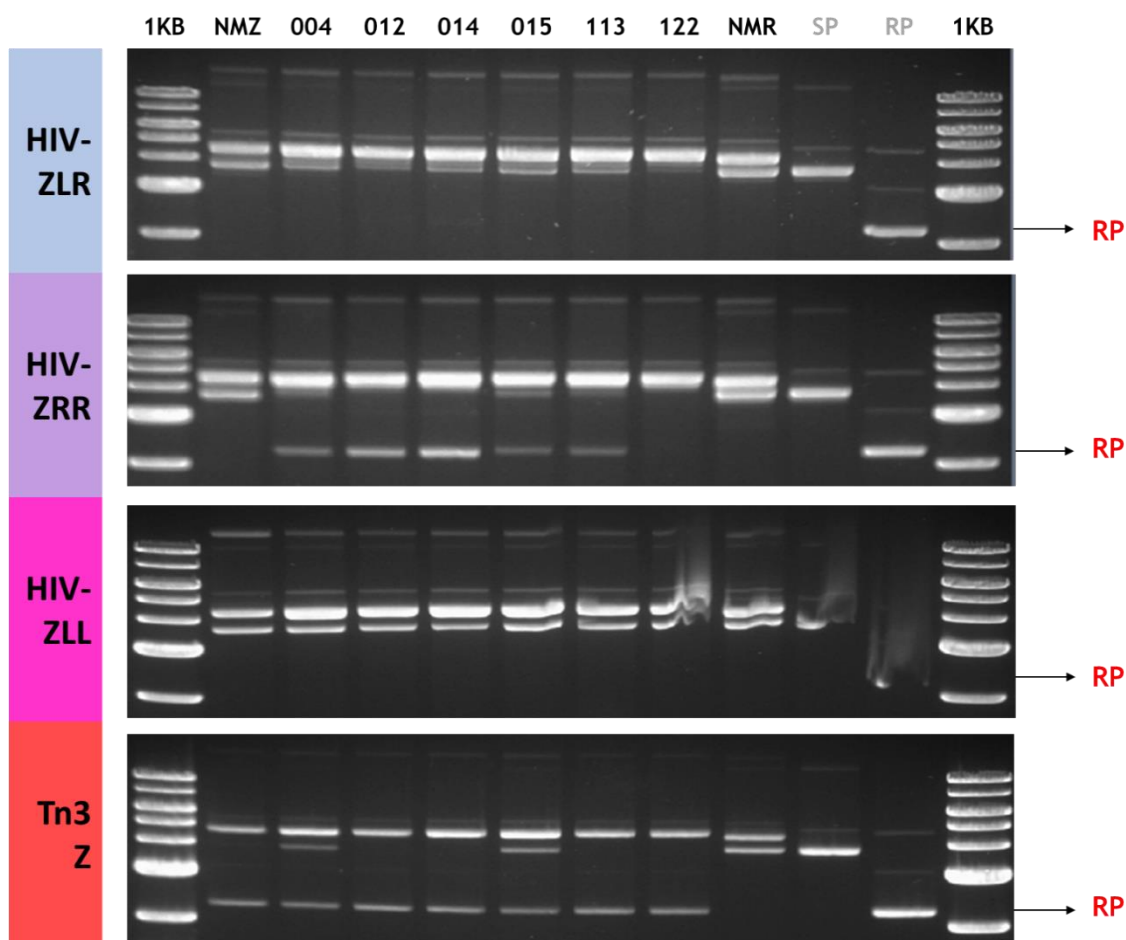


Figure 3.13: Time course of ZR012 activity. A. on pJU002 (HIV-ZRR). B. on pJU003 (HIV-ZLR). To identify the time point of substrate depletion of ZR012 on pJU003, a 12-hour *in vivo* recombination time course was carried out. The expression plasmid of ZR012 (pJUM012) and FO2 (pFO2) were transformed into COMPETENT *E. coli* cells containing either pJU002 or pJU003. After 90 minutes expression, cells were inoculated into L-broth (containing kanamycin and ampicillin). Time points were counted after the initial 90-minute expression step after which, DNA products were recovered from the cultures and run on agarose gels. pFO2 is a standard expression plasmid for NM resolvase; it serves as a negative control as it should not recombine pJU002 or pJU003. Substrate depletion of pJU003 seems to occur from the onset although it does not progress as fast or as completely as on pJU002. Recombination on pJU002 leads to recombination products which can be observed from 10 hours. A smaller band (marked with ?) that appears in the pJU002 gel at 8 hours is unidentified.



NMZ:	Tn3 NM ZFR
ZR004:	NM + I77L R120L E132A I138V
ZR012:	NM + K29E I77L V107F
ZR014:	NM + K29E I77L V107F R120L E132A I138V
ZR015:	NM + K29E I77L R120L E132A I138V
ZR013:	NM + I77L V107F
ZR122:	K29E G70S V107L
NM:	Tn3 NM resolvase

Figure 3.14: Recombination Activity of designed mutants on pJU001, pJU002, pJU003 and pMP53. The expression plasmids of the proteins listed above were transformed into *E. coli* cells containing either pJU001, pJU002, pJU003 or pMP53. After 90 minutes expression, cells were inoculated into L-broth (containing kanamycin and ampicillin). DNA products recovered from overnight cultures were run on agarose gels. On pJU003, there is depletion of the substrate plasmid with several mutant ZFRs. ZR004, ZR012, ZR014, ZR015 and ZR113 show recombination on pJU002. ZR014 shows improved recombination activity on pJU002 over ZR004 and ZR012. None of the ZFRs show recombination on pJU001. All mutant ZFRs still work on Tn3Z substrate, pMP53, but to varying degrees. Tn3 NM ZFR does not recombine pJU001, pJU002 or pJU003 but works efficiently on its own substrate.

3.3.7 Strategy 4: Designing active TATA-CR catalytic domains based on position -3

3.3.7.1 Analysis of CR_TATA left core sequence

The central 16-bp CR_TATA target site is quite symmetrical. The left and right 8 bp half-sites are identical except at position -3. Tn3 *res* site I has a Thymine at position -3 on both half-sites. However, the TATA target sequence has an Adenine at position -3 of the left half-site and a Thymine on the right half-site (Fig. 3.15). The left half-site of the full HIV Z-site has posed considerable challenges to the selection of active Tn3 NM mutant ZFRs with limited activity (cleavage) on the HIV-ZLR substrate plasmid and no visible activity on HIV-ZLL. It could be inferred from this that the nucleotide change(s) at position -3 is responsible for loss of recombination activity.

	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8
Tn3:	A	A	A	T	A	T	T	A	T	A	A	A	T	T	A	T
	T	T	T	A	T	A	A	T	A	T	T	T	A	A	T	A
HIV:	G	C	T	G	C	A	T	A	T	A	A	G	C	A	G	C
-ZLR	C	G	A	C	G	T	A	T	A	T	T	C	G	T	C	G
HIV:	G	C	T	G	C	A	T	A	T	A	T	G	C	A	G	C
-ZLL	C	G	A	C	G	T	A	T	A	T	A	C	G	T	C	G
HIV:	G	C	T	G	C	T	T	A	T	A	A	G	C	A	G	C
-ZRR	C	G	A	C	G	A	A	T	A	T	T	C	G	T	C	G

Figure 3.15: The nucleotide at position -3 of the Z-site is significant. The HIV recombinase domain target sequence (HIV-ZLR) is quite symmetrical. The left and right 8 bp half-sites only differ at position -3. The central 16-bp of Tn3 *res* site has a T at position -3 of both half-sites. The arrangement in HIV-ZRR is similar. However, in HIV-ZLL and HIV-ZLR, the nucleotides at these positions differ from the canonical *res* nucleotides. Since HIV-ZLR and HIV-ZRR only differ at position -3, any change in the activity of a ZFR on these two Z-sites would stem from this difference.

3.3.7.2 *Analysing position -3 nucleotide preference of Tn3 Z-site*

To confirm that the nucleotide change(s) at position -3 is responsible for loss of recombination activity by mutant ZFRs on HIV-ZLL and HIV-ZLR, the recombination activity of Tn3 NM ZFR was characterized on Tn3 Z-substrates with all 4 possible nucleotides at position -3. To ensure that the cause of any observed change in activity relative to that on the wild-type Z-site is due to the change of the nucleotide at position -3, and not just the context of the nucleotide's position in relation to its flanking sequence, the Z-sites were designed as left and right palindromes of the native Tn3 Z-site, resulting in 8 different Z-sites (Fig. 3.16).

The left palindromic Z-sites were called Tn3LP, Tn3LPA, Tn3LPC and Tn3LPG with the last letter defining what nucleotide position -3 was changed to. The right palindromes were similarly named Tn3RP, Tn3RPA, Tn3RPC and Tn3RPG. The substrate plasmids were designed as before, with two Z-sites in direct repeat flanking the *galK* gene. The eight Z-substrate plasmids were called pJTn3LP, pJTn3LPA, pJTn3LPC, pJTn3LPG, pJTn3RP, pJTn3RPA, pJTn3RPC and pJTn3RPG based on their respective sites. The native Tn3 Z-substrate plasmid (pMP53) was used as a positive control for the experiments, and should reflect maximum recombination activity.

Tn3 NM ZFR showed the same level of recombination activity on pJTn3LP and pJTn3RP as on pMP53, implying that recombination activity on its palindromic half-site substrates is equally as efficient as on its native site. However, its activity on the substrates with modified position -3 was relatively insignificant; all colonies appeared red on MacConkey agar plates, and very faint bands of resolution products were observed (if any) on agarose gels (Fig. 3.17). It should be noted that the Tn3LPA Z-site (with position -3/3 nucleotides similar to that in HIV-ZLL) actually seems to recombine most of the mutant sites. This confirmed that changing the nucleotide at position -3 from T to any other nucleotide hinders the catalysis of recombination at the Z-sites.

			-3	-2	-1	1	2	3		
Tn3:	GCGTGGGCGAGC	AAATA	T	T	A	T	A	A	ATTAT	AGCCGCCACGC
Tn3LP:		AAATA	T	T	A	T	A	A	TATTT	
Tn3LPA:		AAATA	A	T	A	T	A	T	TATTT	
Tn3LPC:		AAATA	C	T	A	T	A	G	TATTT	
Tn3LPG:		AAATA	G	T	A	T	A	C	TATTT	
Tn3RP:		ATAAT	T	T	A	T	A	A	ATTAT	
Tn3RPA:		ATAAT	A	T	A	T	A	T	ATTAT	
Tn3RPC:		ATAAT	C	T	A	T	A	G	ATTAT	
Tn3RPG:		ATAAT	G	T	A	T	A	C	ATTAT	

Figure 3.16: New position -3 permutation Tn3 Z-sites. Eight new Z-sites having all possible nucleotides (Thymine, Adenine, Cytosine and Guanine) at position -3 of both half-sites of the Tn3 Z-site were constructed. The figure shows the top-strand sequence of the Z-sites. The Tn3 Z-site (Tn3) is shown at the top. The central 16 bp is flanked by 9-bp Zif268 binding sites (GCGTGGGCG) and a 3-bp spacer (AGC). Tn3LP is the palindrome of the left half of the 16-bp core sequence of Tn3 crossover site. Tn3RP is the right palindromic sequence. The naming of the new Z-sites is based on the identity of the nucleotide at position -3 on the top strand of Z-site. For example, Tn3LPA implies a change at position -3 of Tn3LP to A, Tn3LPC to C and Tn3LPG to G.

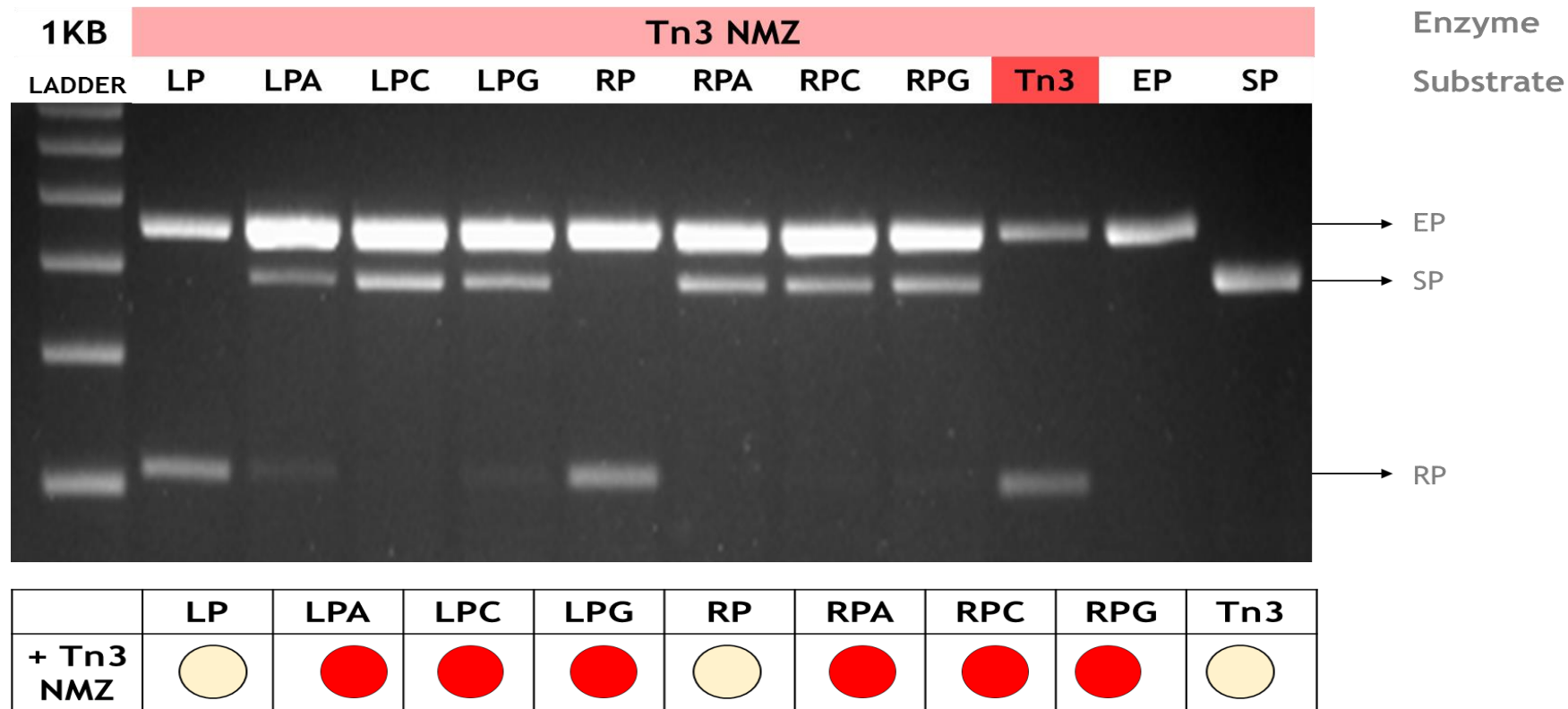


Figure 3.17: Activity of Tn3 NMZ on position -3 permutation substrates. The expression plasmid of Tn3 NMZ (pMP59) was transformed into *E. coli* cells containing pJTn3LP, pJTn3LPA, pJTn3LPC, pJTn3LPG, pJTn3RP, pJTn3RPA, pJTn3RPC, pJTn3RPG, or the Tn3 Z-substrate plasmid (pMP53). After 90 minutes expression, cells were inoculated into L-broth (containing kanamycin and ampicillin). DNA products recovered from overnight cultures were run on agarose gels. A change in the nucleotide at position -3 of Tn3 Z-site critically affects the recombination activity of Tn3 NMZ. NMZ carries out complete resolution on pJTn3LP, pJTn3RP and pMP53 with resolution product bands on agarose gels and white colonies on red plates.

Since position -3 is important at some stage of recombination (sequence recognition, binding, synapsis, cleavage or re-ligation), to re-engineer the Tn3 resolvase catalytic domain for recombination catalysis on non-native sites, it might be essential to decipher the rules for its recognition. The presence of this change in the CR_TATA target makes this imperative here.

3.3.7.3 Library selection for Tn3 ZFR mutants with activity on altered position -3 substrate plasmids

Since nucleotide identity at position -3 is critical to recombination activity, identifying mutants that catalyse recombination on pJTn3LPA and pJTn3RPA might provide sufficient information for the design of mutants that can recombine pJU001 and pJU003. Six available libraries (generated by C. Proudfoot: see Table 3.4) were screened on pJTn3LPA and pJTn3RPA. From thousands of colonies, hundreds of pale colonies were identified, streaked out and analysed; 30 of these were validated hits. Validated hits were colonies that streaked out as white colonies and had the appropriately-sized recombination product band on an agarose gel. Validated hits came mostly from Lib 1 and Lib 5; none came from Lib 3 and Lib 4. The ZFR-encoding plasmids were excised and sent for sequencing. Sequencing resulted in the identification of 16 independent mutants. Analysis of the sequencing data and consensus generation led to the identification of specific mutations that might be responsible for the gain of catalytic activity on these modified Tn3 Z-sites (Fig. 3.18). These mutants were retested on pJTn3LPA and/or pJTn3RPA to confirm recombination activity (Table 3.5). None of the identified mutants were active on pJU001 (HIV-ZLL) or pJU003 (HIV-ZLR) (data not shown). The 6 libraries were also tested HIV-ZLL and HIV-ZLR and no white colonies were observed on any plates.

Table 3.4: Library description

	Library Name	Description
1.	Lib1 (NM + V107F dPTP)	<i>PCR mutagenesis of the full Tn3 NM (+V107F) resolvase catalytic domain range with dPTP in the reaction mixture. The previous name of this library was "pAMC11 dPTP". It was generated on 3/10/2006 from 265,000 clones.</i>
2.	Lib 2 (NM + V107F 8-oxo-G)	<i>PCR mutagenesis of the full Tn3 NM (+V107F) resolvase catalytic domain range with 8-oxo-G in the reaction mixture. The previous name of this library was "pAMC11 8-oxo-G". It was generated on 3/10/2006 from 176,000 clones.</i>
3.	Lib 3 (E-helix dPTP)	<i>PCR mutagenesis of the E-helix region of Tn3 NM resolvase catalytic domain range with dPTP in the reaction mixture. The previous name of this library was "E-helix dPTP". It was generated on 22/05/2009 from 1,300,000 clones.</i>
4.	Lib 4 (E-helix 8-oxo-G)	<i>PCR mutagenesis of the E-helix region of Tn3 NM resolvase catalytic domain range with 8-oxo-G in the reaction mixture. The previous name of this library was "E-helix 8-oxo-G". It was generated on 22/05/2009 from 1,500,000 clones.</i>
5.	Lib 5 (NM + I77L + V107F dPTP)	<i>PCR mutagenesis of the full Tn3 NM (+ I77L + V107F) resolvase catalytic domain range with dPTP in the reaction mixture. The previous name of this library was "I77L dPTP Lib". It was generated on 22/05/2009 from 22,800 clones.</i>
6.	Lib 6 (NM + 177L + V107F 8-oxo-G)	<i>PCR mutagenesis of the full Tn3 NM (+ I77L + V107F) resolvase catalytic domain range with 8-oxo-G in the reaction mixture. The previous name of this library was "I77L 8-oxo-G lib". It was generated on 22/05/2009 from 84,200 clones.</i>

These libraries were designed and constructed by C. Proudfoot.

Of the 16 independent mutants identified, the most active mutants were p3mut3, p3mut7, p3mut8, p3mut12, p3mut13, p3mut14 and p3mut16. It was interesting to note that where tested, the mutants had highest activity on either pJTn3LPA or pJTn3RPA, showing minimal recombination activity on the other if any. Only p3mut3 (same as ZR113: Section 3.3.6) showed the same level of activity on both substrates. p3mut7, p3mut8, p3mut14 and p3mut16 were most active on LPA while p3mut12, and p3mut13 were active on pJTn3RPA (Table 3.5). The single mutant p3mut7 (V107L) recombined pJTn3LPA efficiently but showed no activity on pJTn3RPA. In combination with F83L, V107L (in p3Mut8), some white colonies

appear on pJTn3RPA. The Q116L mutation in p3mut14 (Q13H V107F Q116L) corresponds to the position of the sole activating mutation of Sin resolvase (Q115R). The Q13 mutation here might also be interesting- p3mut15 (I3T Q13L V63G F83L V107F) which shows some activity on pJTn3LPA but not pJTn3RPA also has a mutation at this position.

For recombination of pJTn3RPA, Q13R seems to be the key mutation as it is present in both mutants p3mut12 (Q13R V107F) and (p3mut13 Q13R V107F E128D) that show significant activity on RPA. It is important to note that there was a huge decrease in colony count on MacConkey agar plates with p3mut12 and p3mut13, suggesting an increased cleavage activity.

Table 3.5: Mutants obtained from ZFR library screening on pJTn3LPA and pJTn3RPA and their activities on pJTn3LPA, pJTn3RPA, pJU002 and pJU003

Mutant Name	Mutations (NM+)	From Lib	LPA	RPA	HIV-ZLL	HIV-ZLR	Most active on
p3mut1	V107F (2)	1, 5					-
p3mut2	V107F I138V (2)	1					LPA
p3mut3	I77L V107F (5)	6					Both
p3mut4	K29E V107F	1					LPA
p3mut5	E46K V107F	1		-			
p3mut6	F34L I77L V107F	6	-				
p3mut7	V107L	1					LPA
p3mut8	F83L V107L	5					LPA
p3mut9	D25G V107L (2)	5					LPA
p3mut10	D25G I77L G87S V107F I138T (2)	5					LPA
p3mut11	D25A I77L V107F	6	-				
p3mut12	Q13R V107F	1					RPA
p3mut13	Q13R V107F E128D	1					RPA
p3mut14	Q13H V107F Q116L	1					LPA
p3mut15	I3T Q13L V63G F83L V107F	1					LPA
p3mut16	I77L I80V V107F Q131R (2)	5					LPA

This table shows the results from the retesting of selected mutants on Tn3LPA, Tn3RPA, HIV-ZLL and HIV-ZLR substrate plasmids. The ratio of the pale to red colour in each circle gives an approximation of the proportion of the 'white' to red colonies (See also Table 3.2). The numbers in the brackets indicate the number of sequenced isolates.

*(distinctly reduced number of colonies)

-not tested

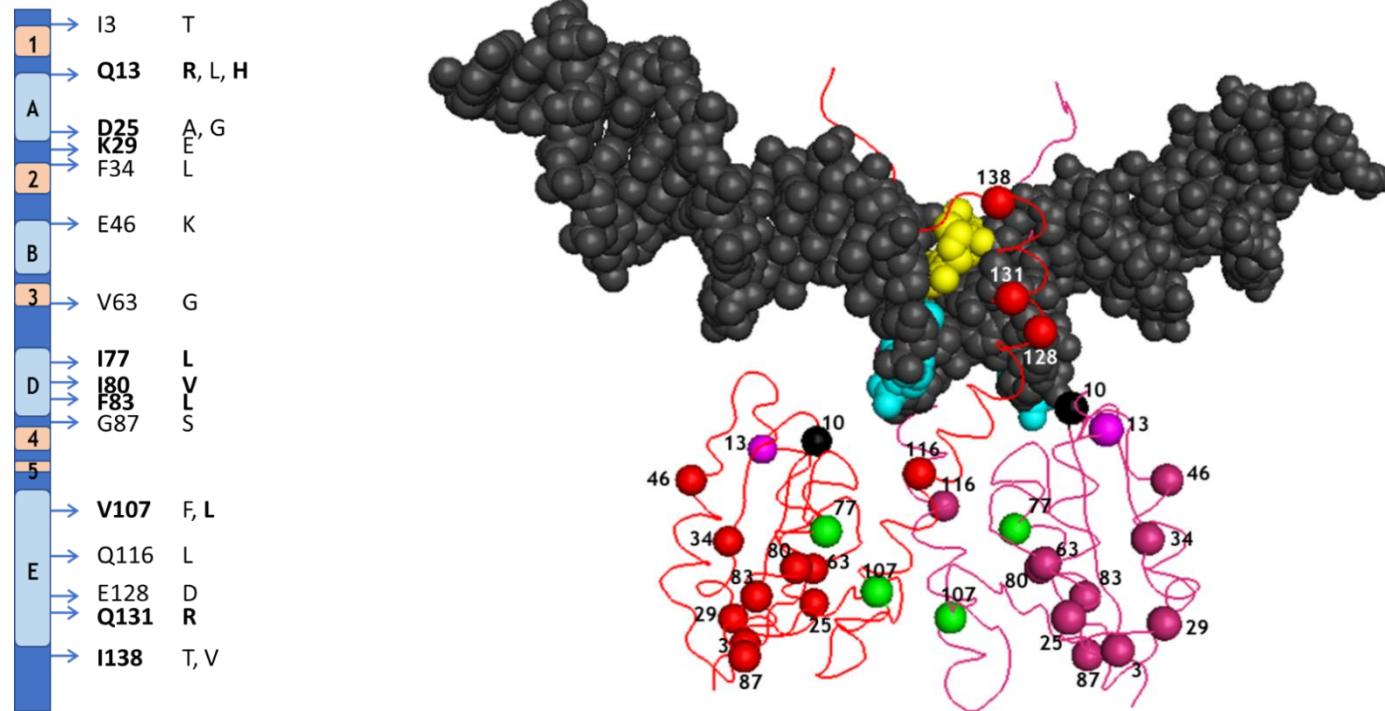


Figure 3.18: Position -3 activating mutations. The identified mutations are shown on a cartoon representation of the crystal structure of 1GDT. As many of the mutants had multiple mutations, some of these residues might not be necessary. DNA is shown as grey spheres in a space-filling model. Thymine at position -3 (Thy16) is shown in yellow and Adenines are shown in cyan. The backbones of monomers A and B are shown in red and pink respectively. S10 is shown as a black sphere; I77 and V107 in green; Q13 in magenta and the remaining mutations according to their monomers. The positions of the residues on the Tn3 resolvase sequence space are shown in a bar on the left with α -helix in blue and β -sheet in peach as predicted by Yang and Steiz, 1995. The activating mutation residues considered significant based on results from *in vivo* recombination assay are in bold.

3.3.8 Rational design of new mutants

Analysis of the mutations revealed a few potentially important residues. Selected mutations were inserted into previous active HIV ZFR mutants (ZR004, ZR012 and ZR014) to generate 10 new mutants (ZFRs 041 to 050) (Table 3.6).

In vivo recombination analysis of these new mutants was carried out on pJU001 and pJU003. This resulted in the identification of a new HIV ZFR mutant with recombination activity on the HIV-ZLR substrate. This new mutant, **ZR045** (NM + Q13R I77L V107F R120L E132A I138V) yields recombination product with pJU003 as substrate; however, on pJU001, there is a depletion of the substrate without appearance of recombination product (Fig. 3.19).

This mutant is simply the original ZR004 (Section 3.3.5; Table 3.3) with V107F and Q13R added. It is worthy of note, though, that this is the first ZFR to show any detectable form of depletion of pJU001 in the recombination assay used. This suggests that the Q13R mutation may have conferred ability for catalytic activity on the substrate Z-site with an A at position -3. The increased substrate depletion observed here with pJU001 and previously with pJTn3LPA and pJTn3RPA (Section 3.3.7.3) suggests that although the ability to cleave position -3-modified Tn3 sites has been introduced by the Q13R mutation, there is potentially more engineering required to get the protein to efficiently religate these sites. Due to time constraints, this was considered beyond the scope of this work.

The product from the recombination activity of ZR045 on pJU003 was excised from the agarose gel, sequence-verified and confirmed as the expected product resulting from the excision of the *galK* gene from pJU003. ZR045 has a V107F mutation which is absent in ZR044. The difference in the activities of ZR044 and ZR045 shows the importance of the V107F mutation in the appearance of recombination products.

Table 3.6: Mutants designed by rational combination of mutations

Active ZFR variants on pJU002	ZR004: I77L R120L E132A I138V ZR113: I77L V107F ZR012: K29E I77L V107F ZR014: K29E I77L V107F R120L E132A I138V
Interesting mutations from library search on pJTn3LPA and pJTn3RPA	V107L (from p3mut7) Q13R (from p3mut12) F83L Q13R V107L (from p3mut8) Q13H V107L Q116L (from p3mut14) Q13H (from p3mut14) I77L I80V V107F Q131R (from p3mut16)
New mutants designed by introducing identified mutations into existing mutants	ZR041: K29E I77L V107L ZR042: I77L V107L R120L E132A I138V ZR043: K29E I77L V107L R120L E132A I138V ZR044: Q13R I77L R120L E132A I138V ZR045: Q13R I77L V107F R120L E132A I138V ZR046: K29E F83L V107L ZR047: K29E F83L V107L R120L E132A I138V ZR048: Q13H I77L V107L Q116L ZR049: Q13H I77L R120L E132A I138V ZR050: K29E I77L I80V V107F Q131R

Selected mutations were cloned from the mutant plasmid vector into the ZFR expression plasmids of available ZFR variants with activity on pJU002. This was done by carefully selecting restriction sites on both plasmids for DNA fragment swapping (Section 2.10.1).

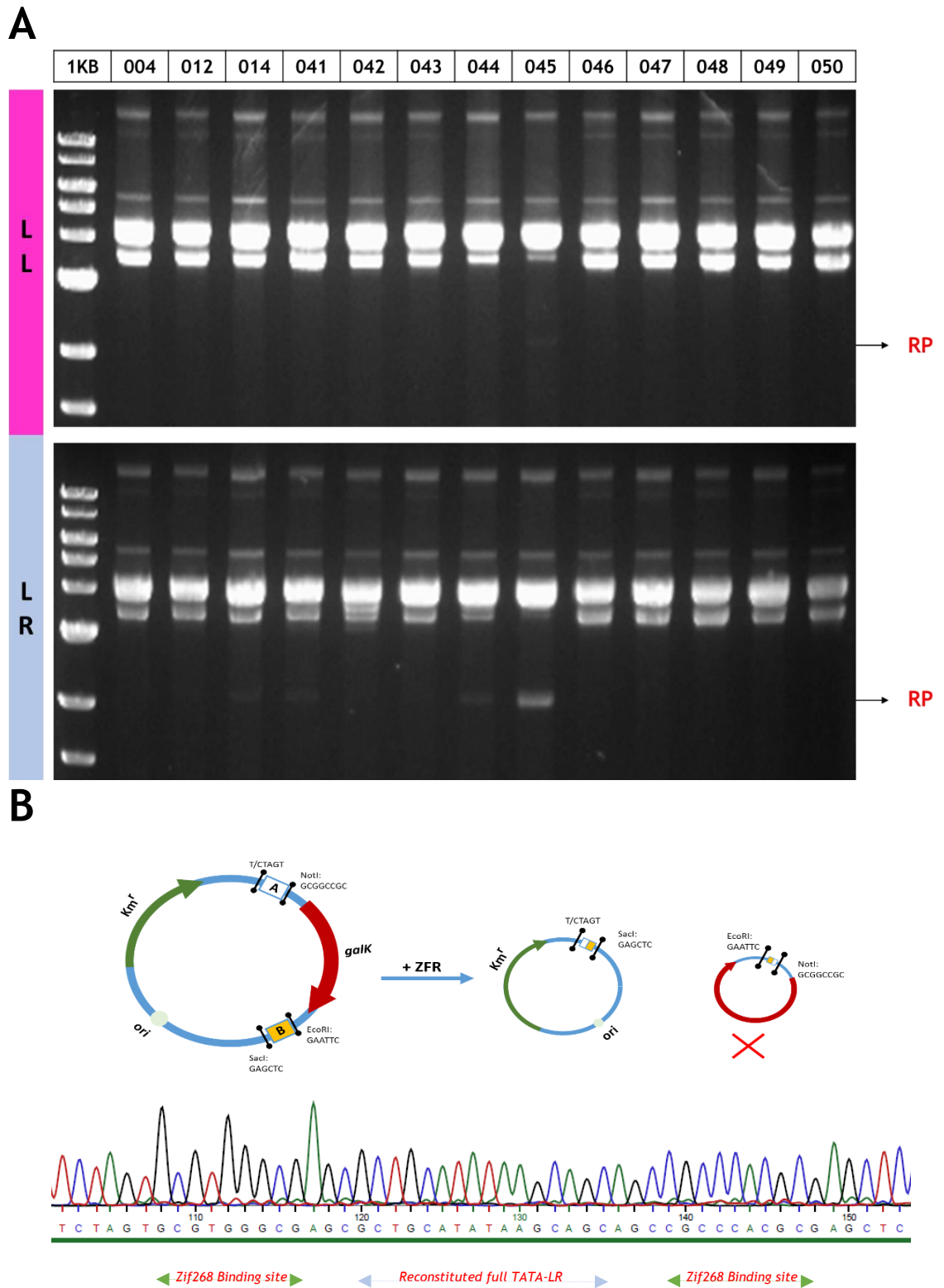


Figure 3.19: Activity of designed mutants on HIV pJU001 and pJU003. ZR045 shows significant recombination activity on pJU003 (HIV-ZLR). Also, for the first time, significant substrate depletion and traces of resolution products are observed on pJU001 (HIV-ZLL) with ZR044 and ZR045; this could imply that the mutations in ZR045 are effective for improved cleavage of Z-sites with non-native position -3 nucleotides. B. Bulk sequencing (Section 2.11) of recombination product of ZR045/ZLR shows expected sequence of product.

3.3.9 Optimizing activity of working mutants

Now that a mutant, ZR045, was obtained with activity on the Zif268-TATA recombinase target site, an attempt to improve its activity was made. Two approaches were taken to do this (Table 3.7). The first involved inserting the K29E mutation (Section 3.3.6) into ZR044 and ZR045 (yielding ZR051 and ZR052 respectively) to check if this mutation increases recombination efficiency. The second approach was to introduce TATA-activating mutations (K29E + V107F) and (Q13R) into the ZR003 (Section 3.3.5) expression plasmid backbone (yielding ZR056 and ZR057 respectively) to check if a different E-helix framework would enhance recombination activity. *In vivo* recombination assays showed that the activities of these mutants do not seem to improve on ZR045 significantly (Fig. 3.20).

Table 3.7: Mutants designed to optimize working mutant activity

Active ZFR variants	ZR044: Q13R I77L R120L E132A I138V ZR045: Q13R I77L V107F R120L E132A I138V ZR003: I77L R120I T126M L135R I138V F140M
Selected mutations for mutant optimization	K29E K29E V107F Q13R
New mutants designed	ZR051: Q13R K29E I77L R120L E132A I138V ZR052: Q13R K29E I77L V107F R120L E132A I138V ZR056: K29E I77L V107F R120I T126M L135R I138V F140M ZR057: Q13R I77L R120I T126M L135R I138V F140M

Selected mutations were cloned from the mutant plasmid vector into the ZFR expression plasmids ZR044, ZR045 and ZR003 (pJUM044, pJUM045 and pJUM003). ZR051 and ZR052 expression plasmids (pJUM051 and pJUM052) were constructed by cloning annealed oligonucleotides (Section 2.7) encoding the K29E mutation (along with Q13R) into pJUM044 and pJUM045. pJUM056 and pJUM057 were cloned by carefully selecting restriction sites for DNA fragment swapping (Section 2.10.1).

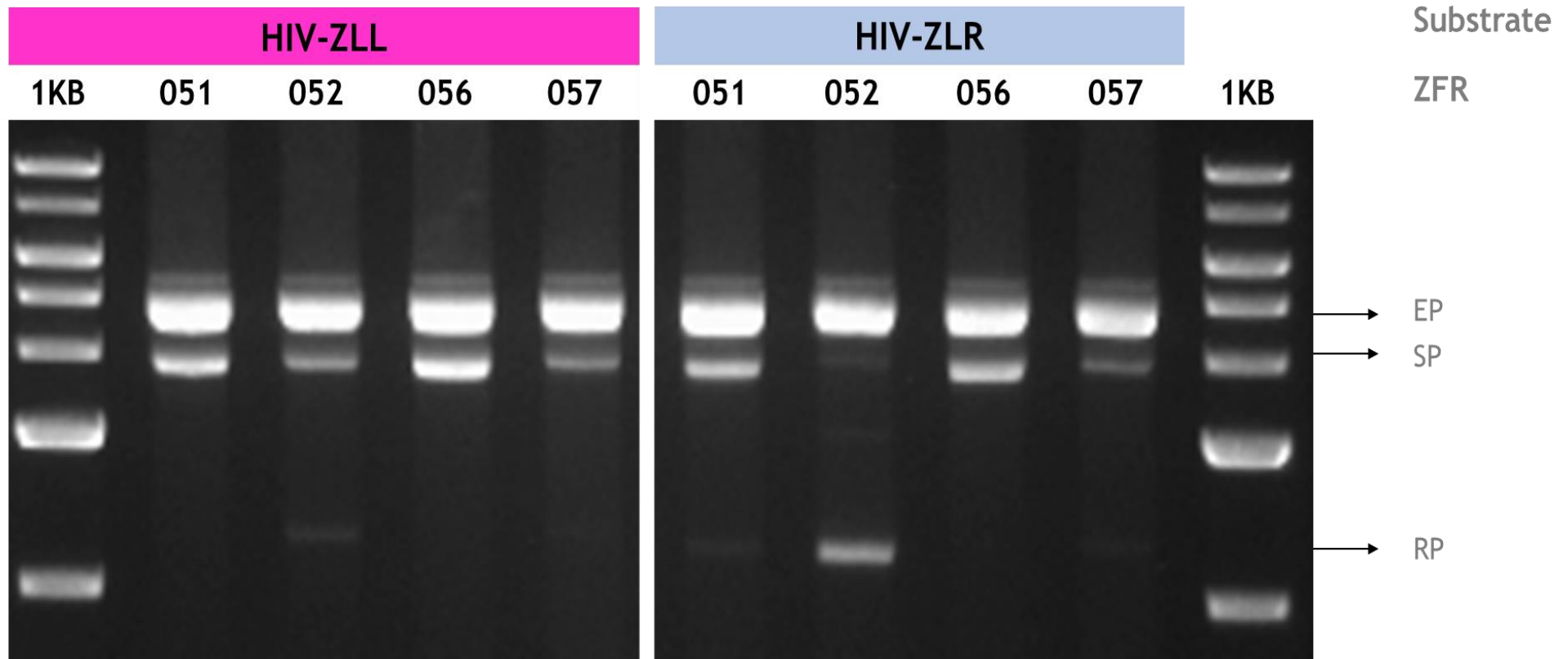


Figure 3.20: Mutant optimization. The expression plasmids of ZR051 (pJUM051), ZR052 (pJUM052), ZR056 (pJUM056) and ZR057 (pJUM057) were transformed into *E. coli* cells containing either pJU001 (HIV-ZLL) or pJU003 (HIV-ZLR). After 90 minutes expression, cells were inoculated into L-broth (containing kanamycin and ampicillin). ZR052 shows minimal recombination product on LL and stable products with LR. ZR056 does not show any activity and the ZR057 mutant shows increased substrate depletion on both substrates.

3.4 Discussion

Here, a target sequence for the CR-based excision of the HIV-1 proviral DNA was carefully selected. By combining rational and random mutagenic approaches, recombinase catalytic domain variants that catalyse recombination at the core 16-bp sequence of the CR_TATA_target site have been engineered. This protein is still being guided by the Zif268 protein to a synthetic target site comprising the Zif268 target site flanking the HIV target central 16 bp. The DNA-binding domain for the native HIV target site needs to be designed and optimised and the subsequent chapters will discuss how this was achieved. A discussion on the biotechnological implications of the design of CRs is provided in Section 6.1.

3.4.1 Determinants of recombination on non-cognate sites

For future work in retargeting the recombinase catalytic domain for genome editing, this work has identified some factors important to consider.

1. Although E-helix mutations allow the retargeting of the recombinase catalytic domain to non-cognate sites, activating mutations in the catalytic region of the protein might improve the activity of the designed/selected mutant.
2. The mutations in this work can be categorised as core catalytic activating mutations (e.g. Q13R and K29E), interface-stabilizing (or destabilizing) mutations (e.g. I77L, V107F and R120L), and E-helix specificity-broadening mutations (e.g. E132A and I138V).
3. The identity of the nucleotides at position -3 of the core 16-bp target sequence is important and could serve as a determining factor for selecting the starting recombinase to engineer for novel specificity.
4. A combination of random and rational engineering approaches is required to target complex sequences.

The central 16-bp sequence of the CR_TATA_target site bears only a 5-bp (from position -2 to +3) similarity to Tn3 *res* site I (Fig. 3.3). Yet, the double mutant ZR113 (NM+ I77L V107F) allows some recombination on the HIV-ZRR substrate and the introduction of the K29E mutation increased this recombination ability

substantially. This infers that the sequence specificity of the catalytic domain is not strict. It has been suggested that an AT-hook motif-like homeodomain, GRRR (residues 141 to 144 of Tn3 resolvase) makes nucleobase-specific interactions with nucleotides 5-8 of the Tn3 *res* site I crossover site (Prorocic, 2009; Coates *et al.*, 2005). This AT-hook motif should discriminate against the TATA_CR_Target sequence in the region (as it is GC-rich) (Fig. 3.3). It is thus possible that it might have taken up new functions in the ZFR context. There is also a left-right asymmetrical behaviour observed here. The helix-turn-helix (HtH) DBD of wild-type Tn3 resolvase binds tightly to the outermost 6 bp of the right half of the target site and so binding at position 5-8 might not be very important while binding at position -3 might take precedence for anchoring the protein (Bednarz, 1990). This could explain the importance of the identity of the nucleotide at position -3 observed in this work. The activity of ZR045 on the HIV-ZLR substrate plasmid (pJU003) might hinge on the presence of the changed nucleotide at position -3 on the left of the crossover site rather than the right. This could explain why significant recombination has still not been observed on the HIV-ZLL substrate (pJU001) although significant cleavage of this substrate plasmid can be seen with ZR045. Sequence discrimination by Tn3 resolvase probably occurs at all stages of recombination and more work will need to be done to identify the residues responsible for directly interacting with position -3 and at what stages of recombination the protein interacts with this position. A nucleotide-residue code might then be built that can simplify the reprogramming of the catalytic domain.

3.4.2 Mutational analysis

The effects of changes at position -3 of the crossover site and of the catalytic domain mutations at different stages of the recombination reaction (binding, synapsis, cleavage, etc) could be studied by comprehensive *in vitro* experiments, but time did not permit this. Such experiments would be required to exhaustively and precisely understand the contribution of each mutation to the activity of ZR045 and other mutants selected in this work. However, some inferences can be made based on the location of the TATA-activating residues in the available crystal structures of $\gamma\delta$ resolvase. Most of the residues maintain a significant distance from the scissile phosphate in two available structures of $\gamma\delta$ resolvase - 1GDT (Yang and Steiz, 1995) and 1ZR4 (Li *et al.*, 2005) (Fig. 3.21). A cartoon

representation of a crystal structure of $\gamma\delta$ resolvase (PDB: 1ZR4 - Li *et al.*, 2005) showing the positions of the activating mutations is provided in Figure 3.22.

Q13R- This is deemed the most important mutation identified in this work as it activated recombination on the full Zif268-TATA target site (HIV-ZLR) in the context of other mutations. It has not been mentioned in previous work, although mutations of Q14 to positively charged residues have been shown to have activating attributes (Olorunniji and Stark, 2009). The proximity of the Q13 residue to the nucleophilic S10 might alter the conformation of the active site allowing the cleavage of a sequence with altered position -3 (Fig. 3.21). Another unique feature of the mutants harbouring the Q13 mutation is the increased substrate depletion observed with them. This might point towards its activating mechanism and further *in vitro* studies would be required to decipher this.

K29E- This residue seems to be out of the active site region of Tn3 resolvase. It is distant from the DNA but is located in a loop right after helix A and might increase or reduce the flexibility of the region influencing the nucleophilic S10 indirectly. A similar mutation, D25G, which was identified in this work and had been previously identified as an activating mutation in Burke *et al.* (2004) might provide a pointer to the effect of K29E. The K29E mutation seems to enhance cleavage as well, as evidenced in the activity of the mutant, ZR122. Its distance from S10 does not change much between the presynaptic and post-cleavage crystal structures. This residue might also be important for 2-3' interface interactions (Section 1.7) and its mutation might stabilize the synaptic tetrameric complexes of the deregulated Tn3 NM mutant further.

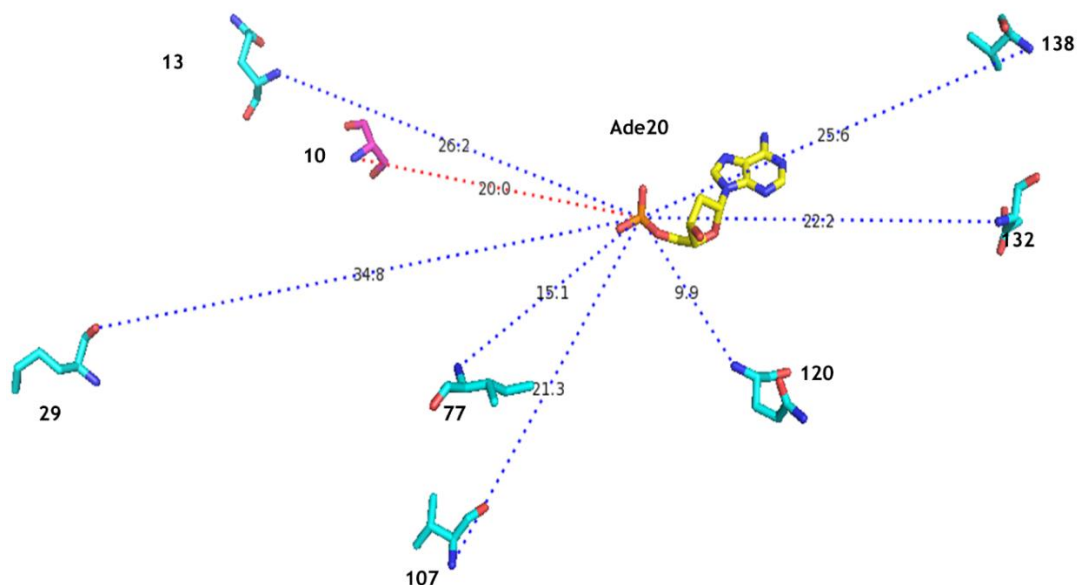
177L and V107F- These two residues seem to stabilize the synaptic interface. Burke *et al.*, (2004) reports multiple interactions of several residues with the I77 residue and with the V107 residue across the two E-helices in the resolvase dimer crystal structure of 1GDT (Yang and Steiz, 1995). These interactions seem to have radically changed in the synaptic tetramer 1ZR4 structure (Li *et al.*, 2005). This shows I77 and V107 in new interactions, with I77 seeming to stabilize antiparallel

helices and V107, parallel helices in the synaptic interface (Fig. 3.23; Fig. 3.24). It is interesting that no mutants that functioned on any of the HIV target substrates (HIV-ZLL, HIV-ZRR and HIV-ZLR) were selected from the E-helix-I77L lib (Section 3.3.5). This points to the significance of the I77L mutation. More *in vitro* studies of these two residues might yield more information about re-targeting the catalytic activity of Tn3 resolvase.

R120L: The side chain interaction between R120 (Q120 in $\gamma\delta$ resolvase) and the activating mutation Y102 on an adjacent monomer is thought to stabilize the synaptic interface through hydrophobic stacking (Li *et al.*, 2005) (Fig. 3.25). It is unclear how this stabilization (or destabilization) of the synaptic interface allows novel site recognition. In the simplest mutant ZR004, reverting the R120L mutation to R120 led to a complete loss of activity. The R120L mutation might allow the ZFR to wrap around the non-cognate target DNA, probably slightly loosely relative to the non-mutated conformation.

E132A and I138V: These two mutations along with the R120L mutation conferred gain-of-function activity on Tn3 NM ZFR to recombine the G-rich Zif268-TATA HIV-ZLR substrate sequence. According to Burke *et al.*, (2004), the side chain of I138 (V138 in $\gamma\delta$ resolvase) contacts the DNA backbone in the minor groove. The I138V mutation was the only activating mutation identified by Burke *et al.*, (2004) in our target E-helix region. Proudfoot *et al.*, (2011) predicted that the E132A mutation might increase the affinity of the protein for DNA in the region due to a decrease in negative charge. It is possible that the E132A and I138V mutations are not specific for our target site and simply broaden the specificity of the protein. However, ZR004 does not recombine the Sin Z-site, implying that the sequence space of this broadening effect is not too wide.

1GDT



1ZR4

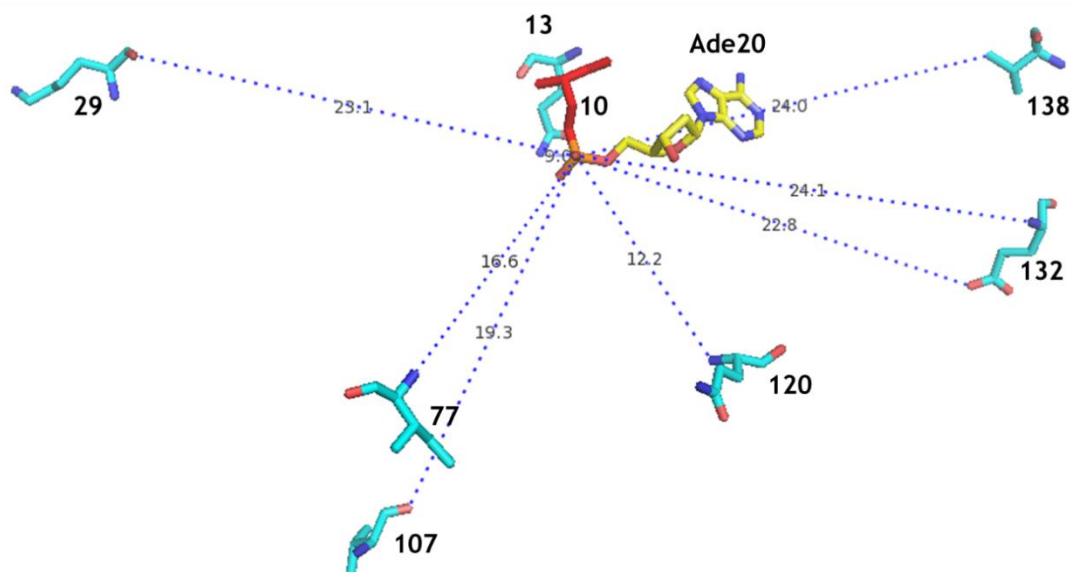


Figure 3.21: Distance from scissile phosphate. Ade20 is shown as a yellow stick, TATA-activating residues as cyan sticks and S10 in pink. Most of the CR_TATA activating mutations stay a considerable distance away from the scissile phosphate in both structures. Q13 due to its proximity to S10 is significantly closer to the DNA backbone but still too far away from Ade20 (and Thy16 or Ade21) to make direct interactions. The distance of Q13 to the scissile phosphate in 1ZR4 (obscured in the image) is 8.0 Angstroms.

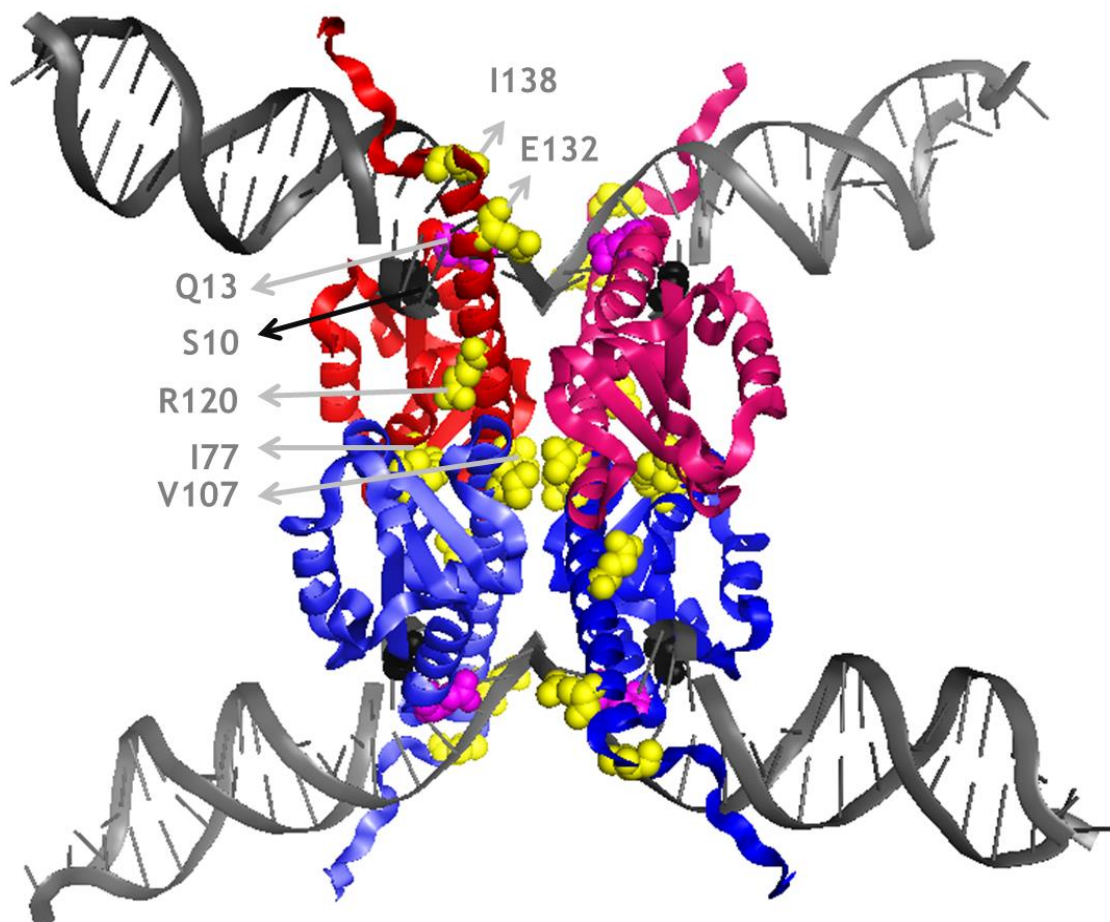


Figure 3.22: Positions of ZR045 activating residues in cartoon representation of $\gamma\delta$ resolvase. Resolvase monomers on the post-cleavage synaptic structure of $\gamma\delta$ resolvase (PDB: 1ZR4 - Li *et al.*, 2005) are shown in red, pink, light blue and deep blue cartoons. Residues are labelled on one monomer to indicate their positions. The activating residues I77, V107, R120 (Q120 in $\gamma\delta$ resolvase), E132 and I138 (V138 in $\gamma\delta$ resolvase) are shown as yellow spheres. Q13 is shown as magenta sphere and S10 as black spheres. DNA is shown as grey cartoon. The E132 and I138 residues are in close proximity to the DNA backbone and might be important for non-direct readout in DNA recognition.

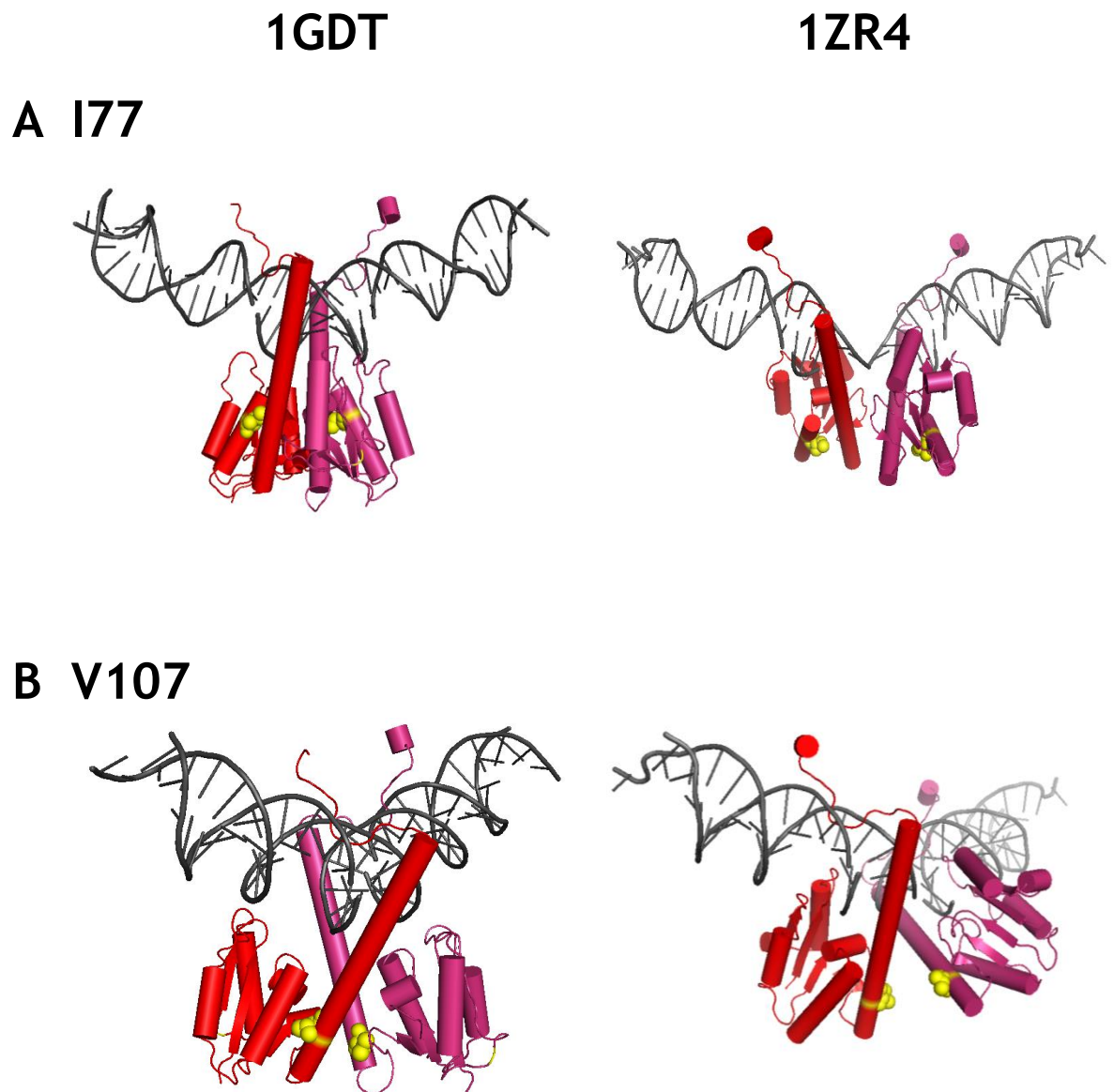
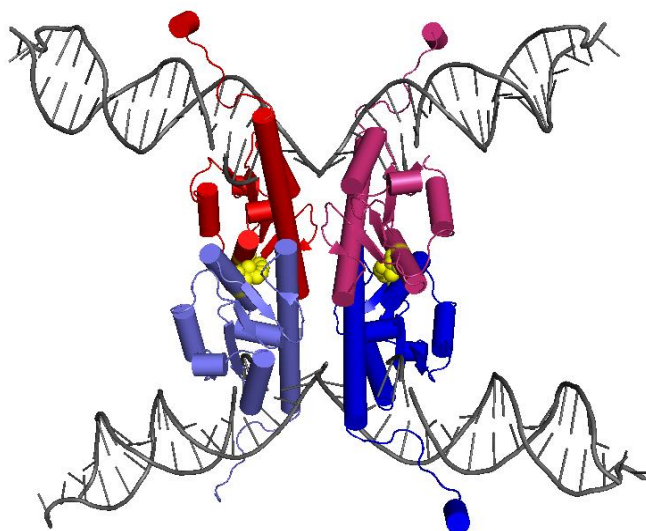


Figure 3.23: Positions of CR_TATA activating residues in cartoon representations of $\gamma\delta$ resolvase - I77 and V107. Monomers of $\gamma\delta$ resolvase in the presynaptic structure (PDB:1GDT- Yang and Steiz, 1995) are shown in red and pink cartoons. The I77 (**A**) and V107 (**B**) residues are shown as yellow spheres. DNA is shown as grey cartoons. **A.** I77 residues interact in *cis* within their own monomers in the presynaptic complex and the conformational changes from pre-synapsis to cleavage expose the residues for interaction with antiparallel helices, stabilizing the synaptic interface. **B.** V107 residues are moved significantly as well and the residues seem to self-interact across parallel monomers (See below).

A 177



B V107

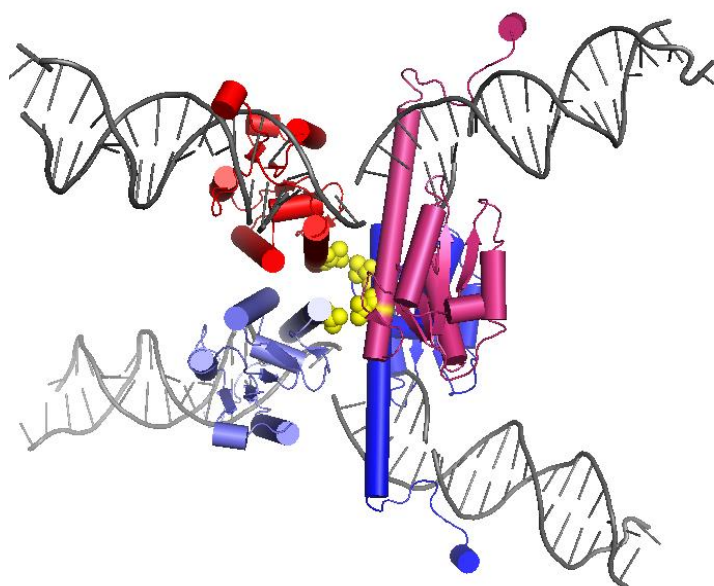


Figure 3.24: Positions of CR_TATA activating residues in tetrameric cartoon representation of $\gamma\delta$ resolvase - I77 and V107. Monomers of $\gamma\delta$ resolvase in the post-cleavage synaptic structure (PDB: 1ZR4 - Li *et al.*, 2005) are shown in red, pink, light blue and deep blue cartoons. The I77 (**A**) and V107 (**B**) residues are shown as yellow spheres. DNA is shown as grey cartoons. **A.** The 1ZR4 structure shows the new interfaces of the I77 interactions. The residues seem to self-interact across anti-parallel monomers. **B.** The full 1ZR4 structure (rotated for a better view) shows the new interfaces of the V107 interactions across parallel monomers, stabilizing the synaptic interface. Phenylalanine instead of valine in this complex might stack better.

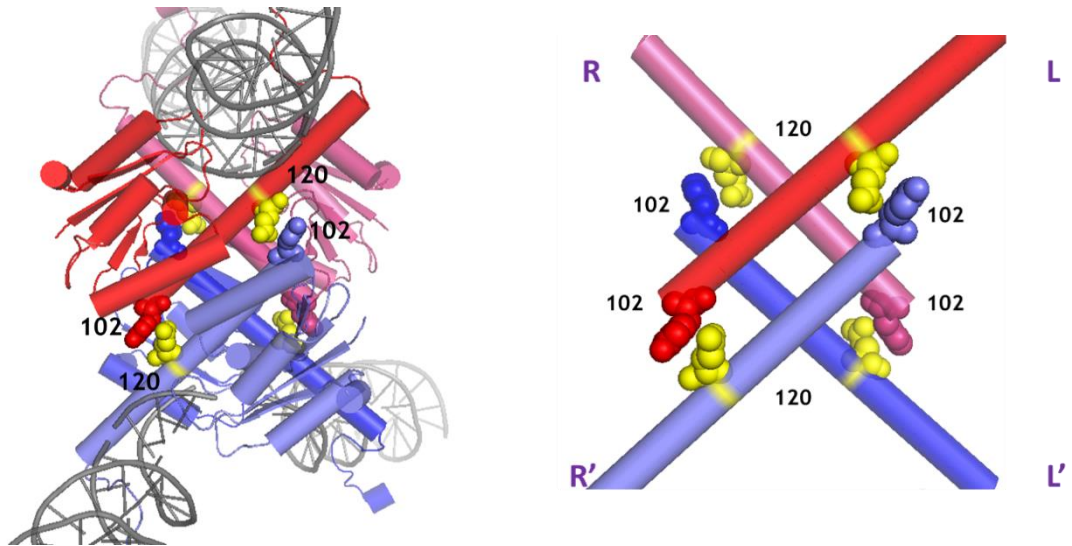


Figure 3.25: Positions of CR_TATA activating residues in cartoon representations of $\gamma\delta$ resolvase - R120. Monomers of $\gamma\delta$ resolvase in the post-cleavage synaptic structure (PDB: 1ZR4 - Li *et al.*, 2005) are shown in red, pink, light blue and deep blue cartoons. The R120 residues (Q120 in $\gamma\delta$ resolvase) are shown as yellow spheres while the Y102 residue is coloured according to the monomer backbone. DNA is shown as grey cartoons. The full 1ZR4 structure shows the interaction of R120 (Q120 in $\gamma\delta$ resolvase) with the activating mutation, Y102. E-helices of 1ZR4 are shown on the right to emphasize these interactions.

Chapter Four: Defining the TALER architecture

4.1 Introduction

This section focuses on the selection and design of the DBDs of the CRs for the excision of HIV-1 proviral DNA. Previous work on engineering CRs have had a strong focus on the use of zinc finger proteins (Table 4.1). As most of these studies have been exploratory in nature, target sequences have been carefully chosen for the simplicity of the creation of the DBD and/or the modified catalytic domains. However, for this work, the sequence constraints discussed in Section 3.3.1 require the careful and strict design of the DBDs to target the sequences flanking the core 16-bp of the CR_TATA_target site. In using ZF proteins for genome editing, it is easier to work forward than backward. That is, it is easier to define genomic target sites based on software-predicted optimal zinc finger binding sequences within a genomic region than it is to design zinc finger proteins for a specific target sequence. The design of ZF proteins for a target sequence can be hampered by the context dependency of adjacent fingers within the protein module, limited zinc finger pools to select from and a natural preference of zinc-finger proteins for G-rich target sequences (Section 1.3.1.1) (Carroll, 2011).

According to Sander *et al.* (2010a), optimal sites for ZF protein-targeting feature more guanines and less thymines, especially at positions 1 and 7 of a three-finger binding site (9 bp) (Fig. 4.1). This finding is a critical limiting factor for the use of zinc finger proteins for generating TATA_CRs. The fitting of the CR_TATA_target site with the 22-bp Z-site architecture shows that at position 7 of the left CR_TATA ZF target site is a thymine (Fig. 4.2). Another significant challenge is that the currently available OPEN zinc finger pool (Section 1.3.1.1) is limited to GNN- and TNN- recognizing finger modules (Maeder *et al.*, 2008) (Fig. 4.1). Where designed and in use, ZF proteins that target CNNs and TNNs have been reported to behave poorly or to have degenerate sequence recognition specificities (Lam *et al.*, 2011). This leaves the remaining triplets within the CR_TATA_target sequence intractable (Fig. 4.2). Altering the 22-bp Z-site architecture to other tolerable Tn3 ZFR Z-site spacer lengths did not eliminate the design challenges (Fig. 4.2). An attempt to use a tool from the Zinc finger consortium, 'Zinc finger targeter' (ZiFiT) to design CR_TATA- targeting zinc finger proteins within the 60-bp CR_TATA_target site based on both the OPEN and CoDA assembly methods

identified no suitable target sites (Sander *et al.*, 2010b; Maeder *et al.*, 2009). This implied that an alternative DBD would have to be used for TATA_CRs. At the inception of this project, the mechanisms, reprogramming and applications of CRISPR-Cas9 genome editing systems were still foundational. So, TALE recombinases were the most suitable candidates to design (Section 1.3.1.2). Since the only target site restriction of TALE DBDs is the requirement of a 5' thymine at position 0, and newer TALE variants with altered and relaxed specificity for this nucleotide have been reported, full programmability for targeting the HIV-1 proviral DNA using TALEs was deemed achievable (Doyle *et al.*, 2013).

Previous work on TALE recombinases has not shown significant recombination activity (Section 1.9.2). It has been predicted that the optimal architecture for TALERS and their T-sites has not yet been defined (Holt, 2014; Mercer *et al.*, 2012). Both of these studies reported work on TALER design focused on two different catalytic domains and two different TALER DBDs. Mercer *et al.* (2012) utilized a mutant Gin invertase catalytic domain fused to N-terminal and C-terminal incremental truncations of wild-type AvrXa7 TALE protein (Section 1.9.2). Holt (2014) focused on Tn3 NM resolvase-based TALERS using N-terminal truncations of a synthetic TALE protein, TALE1297 (from TALEN1297) (Sander *et al.*, 2011). Most TALE proteins, including AvrXa7 and TALE1297 have quite similar sequences and as both findings reported quite different active spacer lengths (32-bp 'gix' spacer for Mercer *et al.*, 2012 and 22-bp/24-bp spacers for Holt, 2014), the Tn3 resolvase-based construct was selected for optimization in this work.

In defining the optimal TALER architecture, the truncation of the N-terminal region of the TALE protein should eliminate unrequired N-terminal domain sequences but retain structural features responsible for TALE binding. Revisiting the TALE structure from Section 1.9.2, there are cryptic repeats in the TALE N-terminal region that contain residues, such as W232, that have been implicated in target site recognition and binding activity of TALE proteins (Cuculis *et al.*, 2015). Although two TALE scaffolds ($\Delta 136/+63$ and $\Delta 153/+47$) are utilized in the generation of active TALEN proteins for genome editing, the importance of the mobility of the resolvase catalytic domain for crossover-site recognition,

staggered DNA cleavage, subunit rotation and cleaved-site re-ligation within a synaptic complex call for a more stringent fusion design. There should be no interference from the TALE protein into the synaptic complex that can destabilize the intermediate stages of recombination. The two domains should be joined together with a linker of appropriate length that allows for optimal modular functionality of DNA-binding by the TALE part and recombination catalysis by the resolvase catalytic domain part.

According to Holt (2014), the four TALERs generated from the fusion of Tn3 NM resolvase (cut-off at residue 144) and N-terminal truncations of TALE1297 ($\Delta 48$, $\Delta 84$, $\Delta 119$, and $\Delta 149$) using a flexible 6aa linker (GSGGSG) showed no significant recombination activity *in vivo* and *in vitro*. The truncation points, ranging about 30 - 36 bp apart, were meant to mimic single 'repeat' increments at each point. Before generating truncation variants of the designed TALERs, Holt (2014) designed a full-length TALER (TALER1) as illustrated in Figure 4.3. The $\Delta 149$ TALE truncation variant from Holt (2014) has been re-numbered to $\Delta 148$ in this work to reflect that 148 residues have been removed from the N-terminal region of the TALE DBD based on numbering by Mak *et al.*, (2012). Of the four different substrate plasmids carrying T-sites of varying spacer lengths (22, 24, 26 and 28 bp), some cleavage activity was observed *in vitro* on T-site22 and T-site24 with the $\Delta 148$ truncation variant (Fig. 4.4). *In vitro* protein characterization allows for the exhaustive analysis of protein activity without the background noise of the cellular environment. It also enables the detection of subtle changes in protein activity that might provide information about its optimization. Since TALER proteins to date have not yielded significant activity in *E. coli* or mammalian cells, a systematic attempt to improve the Tn3-based TALER architecture, define optimal T-sites and characterize Tn3-TALER properties was made here.

Table 4.1: Published work on CRs

	Published work	Catalytic domain	DNA-binding domain	Target cells
1.	Akopian <i>et al.</i> , 2003	Tn3 resolvase	ZF protein	<i>E. coli</i>
2.	Gordley <i>et al.</i> , 2007	Gin invertase	ZF protein	<i>E. coli</i> , HEK293
3.	Gordley <i>et al.</i> , 2009	Gin invertase	ZF protein	HEK293
4.	Gersbach <i>et al.</i> , 2010	Gin invertase	ZF protein	<i>E. coli</i>
5.	Gersbach <i>et al.</i> , 2011	Gin invertase	ZF protein	HEK293, HeLa, HuH-7, NIH3T3
6.	Gaj <i>et al.</i> , 2011	Gin invertase, Tn3 resolvase	ZF protein	HEK293
7.	Proudfoot <i>et al.</i> , 2011	Tn3 resolvase	ZF protein	<i>E. coli</i>
8.	Prorocic <i>et al.</i> , 2011	Tn3 resolvase	ZF protein	<i>E. coli</i> , <i>in vitro</i> *
9.	Mercer <i>et al.</i> , 2012	Gin invertase	TALE protein	<i>E. coli</i> , HEK293
10.	Gaj <i>et al.</i> , 2013	Gin invertase	ZF protein	HEK293
11.	Gaj <i>et al.</i> , 2014	Gin invertase	ZF protein	HEK293
12.	Sirk <i>et al.</i> , 2014	β recombinase, Sin resolvase	ZF protein	<i>E. coli</i>
13.	Holt, 2014 [†]	Tn3 resolvase	TALE protein	<i>In vitro</i> *
14.	Chaikind <i>et al.</i> , 2016	Gin invertase	CRISPR-dCas9	HEK293

**in vitro* analysis requires *E. coli* cells only for recombinant protein expression

[†]PhD thesis

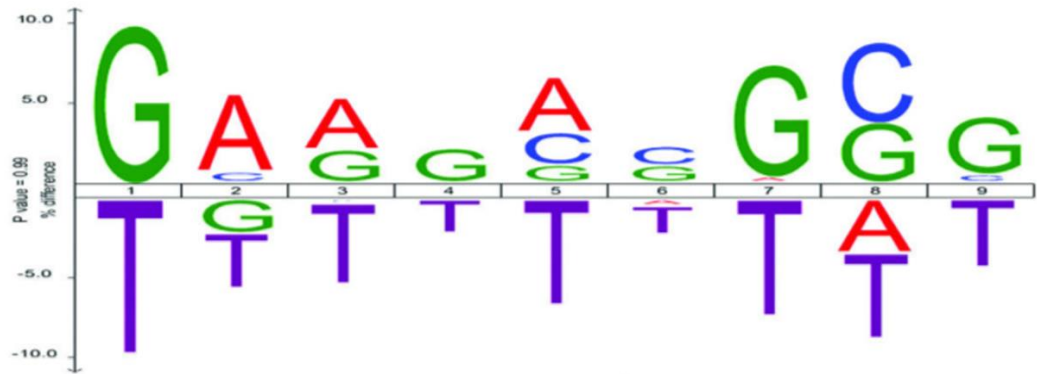
HEK293: Human embryonic kidney cells (in some cases, variants of HEK293 such as the HEK293T mutant cells were used)

HeLa: Human epithelial cervical cancer cells

HuH-7: Human hepatocarcinoma cells

NIH3T3: NIH 3T3 mouse embryonic fibroblasts.

A



B

F1				F2				F3			
AAA	ACA	AGA	ATA	AAA	ACA	AGA	ATA	AAA	ACA	AGA	ATA
AAC	ACC	AGC	ATC	AAC	ACC	AGC	ATC	AAC	ACC	AGC	ATC
AAG	ACG	AGG	ATG	AAG	ACG	AGG	ATG	AAG	ACG	AGG	ATG
AAT	ACT	AGT	ATT	AAT	ACT	AGT	ATT	AAT	ACT	AGT	ATT
CAA	CCA	CGA	CTA	CAA	CCA	CGA	CTA	CAA	CCA	CGA	CTA
CAC	CCC	CGC	CTC	CAC	CCC	CGC	CTC	CAC	CCC	CGC	CTC
CAG	CCG	CGG	CTG	CAG	CCG	CGG	CTG	CAG	CCG	CGG	CTG
CAT	CCT	CGT	CTT	CAT	CCT	CGT	CTT	CAT	CCT	CGT	CTT
GAA	GCA	GGA	GTA	GAA	GCA	GGA	GTA	GAA	GCA	GGA	GTA
GAC	GCC	GGC	GTC	GAC	GCC	GGC	GTC	GAC	GCC	GGC	GTC
GAG	GCG	GGG	GTG	GAG	GCG	GGG	GTG	GAG	GCG	GGG	GTG
GAT	GCT	GGT	GTT	GAT	GCT	GGT	GTT	GAT	GCT	GGT	GTT
TAA	TCA	TGA	TTA	TAA	TCA	TGA	TTA	TAA	TCA	TGA	TTA
TAC	TCC	TGC	TTC	TAC	TCC	TGC	TTC	TAC	TCC	TGC	TTC
TAG	TCG	TGG	TTG	TAG	TCG	TGG	TTG	TAG	TCG	TGG	TTG
TAT	TCT	TGT	TTT	TAT	TCT	TGT	TTT	TAT	TCT	TGT	TTT

Figure 4.1: The limits of zinc finger targeting. A. Icelogo showing zinc finger nucleotide preference at positions 1 to 9 (5' to 3') of a 3-finger target site (from Sander *et al.*, 2010a). A pool containing 135 9-bp zinc finger target sites for which the OPEN pool did (active) or did not (inactive) generate ZF proteins that yielded significant activity on a bacterial two-hybrid (B2H) reporter assay was analysed. The height of the nucleotide in the logo indicates the percentage difference in the composition of the nucleotide at that position in the 'active' dataset and the complete data pool analysed. Although the pool analysed in this assay was limited to GNN and TNN fingers, a strong preference for Gs over Ts is shown across the 9 nucleotide positions especially at positions 1, 5 and 7. B. The current targeting capacity of the OPEN zinc finger pool. Finger availability in the pool for triplet nucleotides is indicated by green shading. It is clear that sequences with ANN and CNN are currently intractable by this pool. Finger availability in the pool does not guarantee its activity (adapted from ZiFit.partners.org; Sander *et al.*, 2010b).

A GGGGAGTGGCCAACCCTCAGAT**GCTGCATATAAGCAGC**TGCTTTTCGCCTGTACTGGGTC
 CCCCTCACCGGTTGGGAGTCTA**CGACGTATATTCGTCG**ACGAAAAGCGGACATGACCCAG

B **HIV-TATA Z-site (Z-22)**

CAACCCTCAGAT**GCTGCATATAAGCAGC**TGC**TTTTTCGCCT**
 ATTGGGAGTCTA**CGACGTATATTCGTCG**ACG**AAAAGCGGA**

Targets:

	F1	F2	F3
Left:	CAA	CCC	TCA
Right:	AGG	CGA	AAA

HIV-TATA Z-site (Z-24)

CCAACCCTCAGAT**GCTGCATATAAGCAGC**TGCT**TTTTTCGCCTG**
 GGTTGGGAGTCTA**CGACGTATATTCGTCG**ACGA**AAAGCGGAC**

Targets:

	F1	F2	F3
Left:	CCA	ACC	CTC
Right:	CAG	GCG	AAA

Figure 4.2: Predicted Z-sites of the HIV CR_TATA_target sequence. A. Full 60-bp CR_TATA_target sequence with the central 16-bp catalytic domain target site in bold. B. The target triplet nucleotides in both left and right ZF binding sites of CR_TATA_target are predominantly CNNs and ANNs. Selection of active ZF proteins on such sites is currently limited.

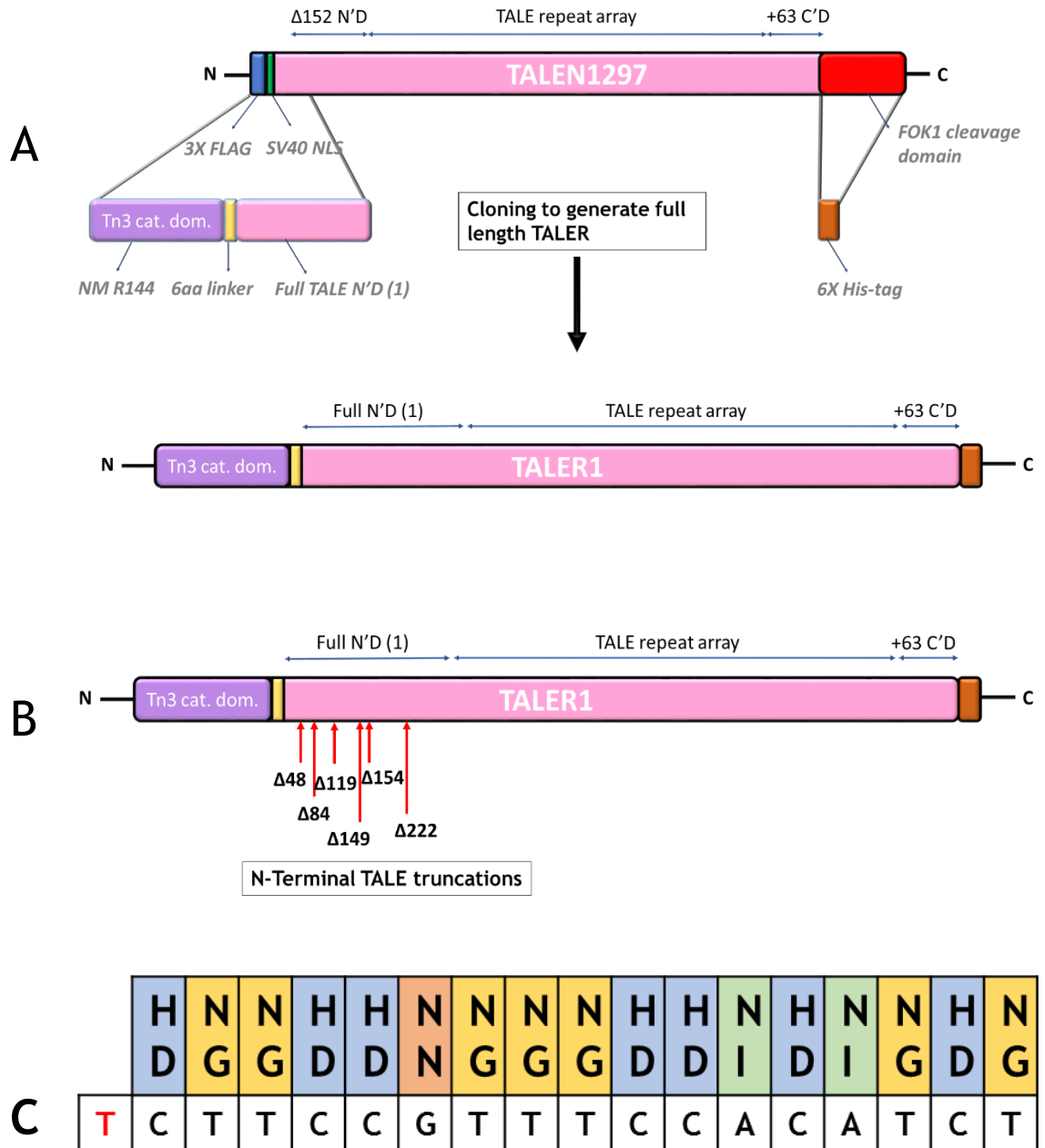


Figure 4.3: Schematic of TALER design. **A.** Holt (2014) generated TALER1 from TALEN1297 (Sander *et al.* 2011) by replacing the N-terminal region with Tn3 NM resolvase catalytic domain (cut-off at residue 144) followed by a 6-aa GSGGSG linker and a full N-terminal sequence of TALE protein with silent restriction sites. The Fok1 nuclease domain was also replaced by a C-terminal 6X histidine tag to enable protein purification. **B.** Several truncation variants of TALER1 were also generated in by Holt (2014) although some of them were simply intermediate cloning points and so were not characterized in the work. The main truncations reported on are $\Delta 48$, $\Delta 84$, $\Delta 119$, and $\Delta 148$. **C.** The 17-bp target sequence of TALE1297 is shown with the respective RVDs. In the design of TALE1297, the RVD HD was chosen for cytosine, NG for thymine, NN for guanine and NI for adenine. The 5' thymine is shown in red and the sequence reads from 5' to 3'.

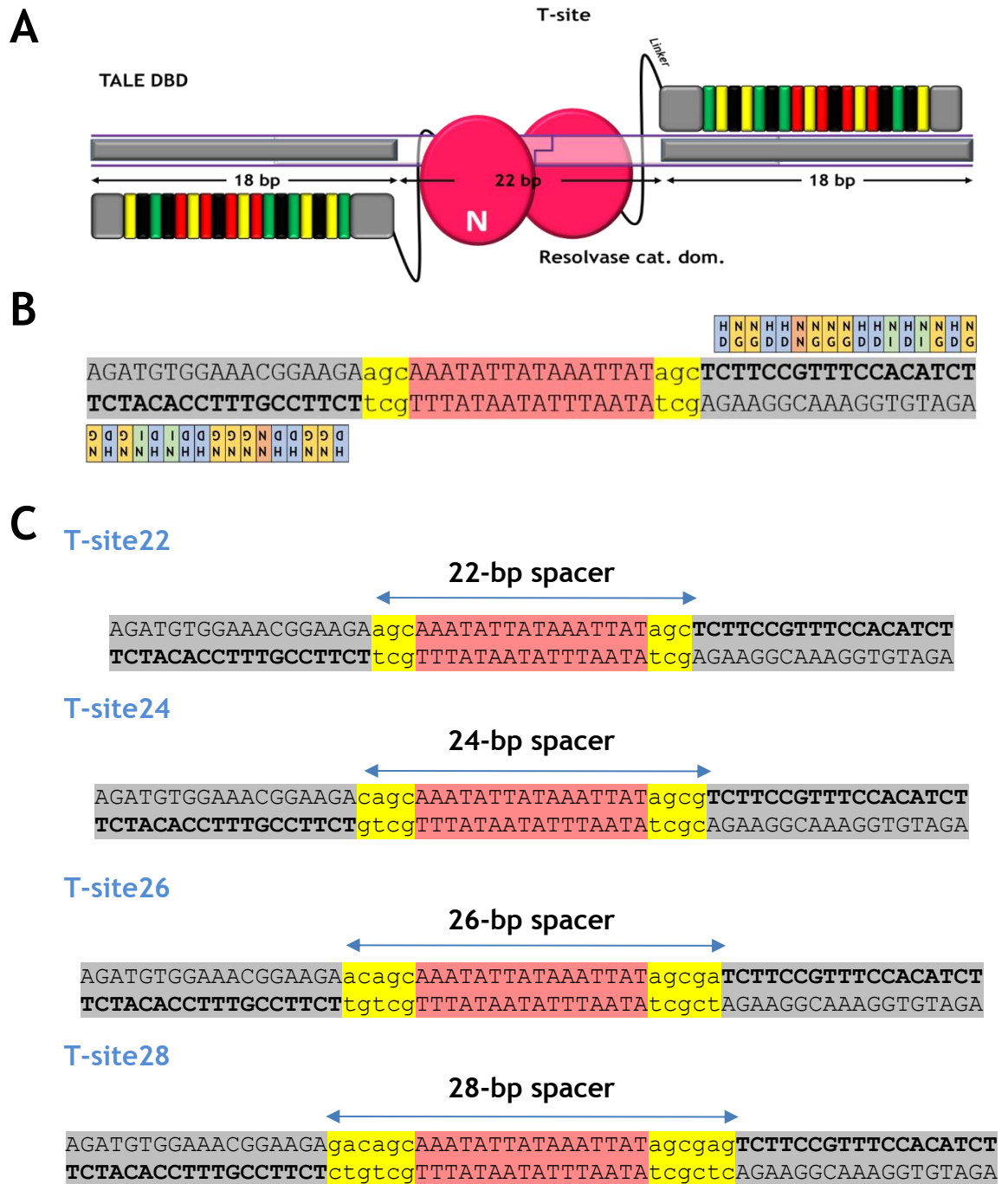


Figure 4.4: TALER target site (T-site). **A.** The optimal architecture for the positioning of TALER at its target site is shown. The recognition sequence of left TALE DBD is on the bottom strand of the T-site in the 5' to 3' direction and of the right DBD on the top strand. **B.** Sequence representation of cartoon design in A is shown. TALE1297 target sites (grey) flank the Tn3 core 16 bp crossover site (red) with 'agc' spacers (yellow) on both sides. TALE DBD target sites are in capitals. **C.** The four T-sites as designed by Holt (2014) are shown. These are the same sequences used in this work. Substrate plasmids have two T-sites in direct repeat flanking a *galk* gene in the low copy number substrate plasmids (TS) and a kanamycin resistance gene in the high copy number substrate plasmids (IVTS).

4.2 Results

4.2.1 *In vivo* properties of TALER

Eleven available TALER variants from Holt (2014) were tested in a MacConkey-based *in vivo* recombination assay (Section 2.11). The expression plasmids for some of these were generated at intermediate cloning steps during the creation of TALER1 and the truncated TALER variants. They all have a +63 C-terminal TALE architecture and are fused to Tn3 NM resolvase (cut-off at residue 144) using different linker lengths and at different truncation points (Fig. 4.3). The four substrate plasmids tested are TS22 (TALER substrate plasmid with 22-bp spaced T-site), TS24, TS26 and TS28 (Fig. 4.3). The spacers consist of the central 16 bp of the Tn3 *res* site I flanked by additional nucleotides symmetrically to make 22-bp, 24-bp, 26-bp and 28-bp sites. The spacer is flanked on both sides by 18-bp TALE binding sites (Fig. 4.4) to make the full-length T-site. Substrate plasmids are as described in Section 2.10.2 with two T-sites in direct repeat flanking a *galk* gene.

The analysis of the bulk DNA recovered from recombination assay plates (section 2.11) did not show any recombination products on agarose gels (Fig. 5). However, there were some white colonies (1% of total colony count) on the MacConkey agar plates, especially with TS22. The analysis of DNA from single white colonies showed the presence of resolution products (data not shown). The presence of long inverted repeats at the T-sites could lead to the formation of cruciform DNA, secondary structures that cause plasmid instability and can promote non-resolvase-mediated excision. Sequence analysis of the recombination products would be required to confirm whether these products stem from TALER-mediated excision.

Tuning *in vivo* protein expression by trying a number of ribosome binding sites (RBS) and different *E. coli* strains did not improve the activity of the proteins. There was still no resolution product observed on agarose gel with the analysis of the bulk DNA (data not shown).

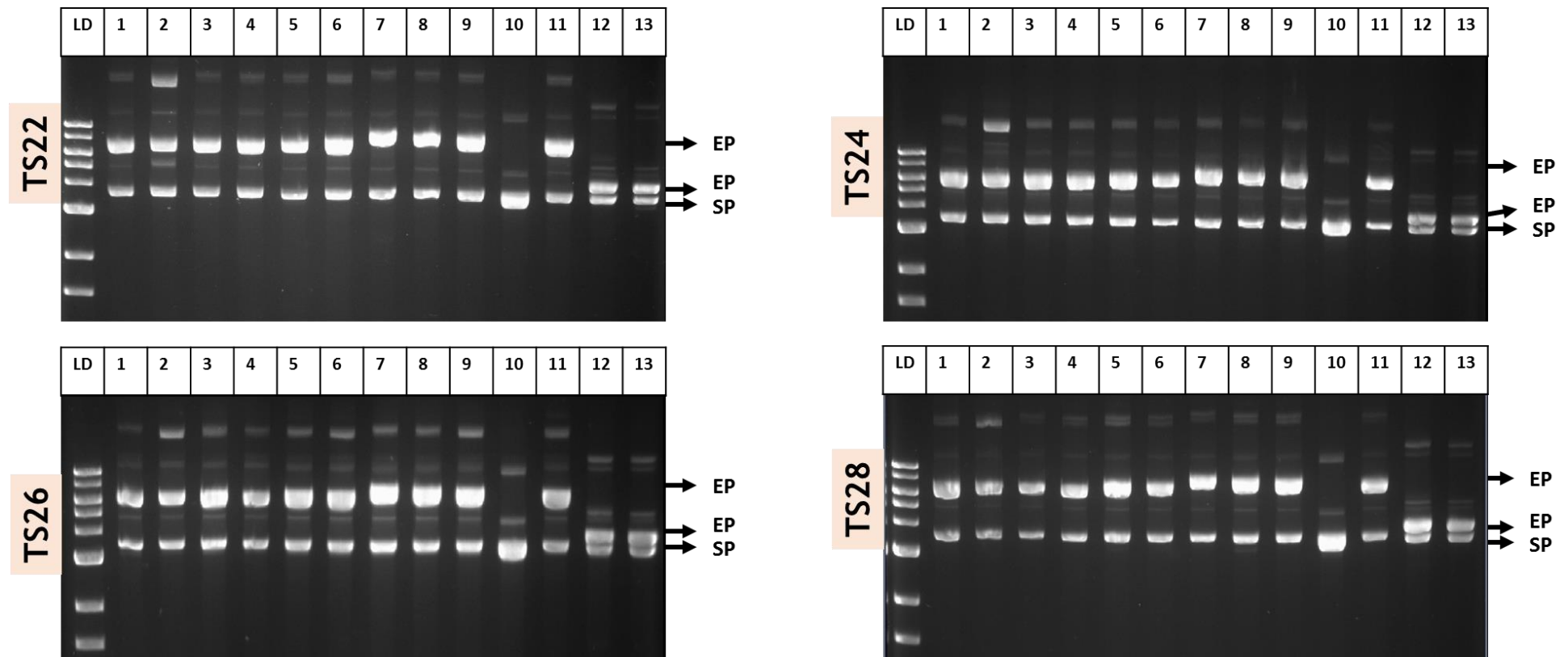


Figure 4.5: *In vivo* activities of TALERs. Analysis of DNA from *in vivo* recombination assay show that all TALER variants analysed do not demonstrate observable excision activity on the substrate plasmids (TS22, TS24, TS26, TS28). 10 turned out to be an incorrect plasmid and shows a different migration pattern on the gel. 1. $\Delta 221$ (+ 6aa linker) 2. $\Delta 221$ (+ 9aa linker) 3. $\Delta 221$ (+ 12aa linker) 4. $\Delta 221$ (+ 7aa linker) 5. $\Delta 153$ 6. $\Delta 153$ (+ 6aa GS linker) 7. TALER1 8. $\Delta D48$ (+ 6aa linker) 9. $\Delta 84$ (+ 6aa linker) 10. $\Delta 119$ (+ 6aa linker) 11. $\Delta 148$ (+ 6aa linker) 12. FO2 (vector control) 13. FO32 (vector control).

4.2.2 *In vitro* structural optimization of TALER activity

Although no recombination activity was observed by Holt (2014) using several truncated TALER variants, it was indicated that the $\Delta 148$ truncation showed what appeared to be a synaptic complex on an electrophoretic mobility shift polyacrylamide gel. It was then inferred that optimizing this truncation variant might improve on its observed cleavage activities. To do this, flexible linkers of lengths varying from 0 to 12 amino acids were designed to fuse the truncated TALE DBD to the Tn3 NM resolvase catalytic domain. The architectures of $\Delta 148$ TALER and the designed linkers are shown in Fig. 4.6. The sequence of TALER1 and the indication of the different truncation points used in this work are provided in Figure 4.7. To improve part complementarity across the ZFRs and TALERs used in this work, the resolvase catalytic domain cut-off point which in Holt (2014) is at residue 144 was shifted 4 residues forward to 148.

Six TALER proteins with the $\Delta 148$ TALE N-terminal domain truncation were expressed and purified using nickel-affinity chromatography as described in Section 2.12. A 3-hour induction at 37 °C yielded significantly better TALER expression than overnight induction at 22 °C (Section 2.12). It is also important to note here that most of the TALER proteins are usually in the insoluble fraction during purification, implying that TALERs may be trapped in inclusion bodies during expression. However, the intricacies of TAL effector protein folding necessitate that denaturing protocols should not be applied in their extraction. Five of the six proteins eluted with a similar chromatographic profile and only TALERO seemed to be very minimally available in the eluate and is predicted to be the most insoluble. Protein concentrations at volumes sufficient for *in vitro* activity characterization were recovered using native purification methods (Section 2.12).

In vitro recombination activity of the 6 proteins was characterized on four TALER substrate plasmids with varying spacer lengths. The four substrate plasmids are called IVTS22 (*in vitro* T-substrate with 22-bp spaced T-site), IVTS24, IVTS26 and IVTS28. *In vitro* recombination was carried out as described in Section 2.16.

Variations in reaction times and protein/DNA concentrations are used across this section to detect subtle differences in protein behaviour, especially where long reaction times seem to show similar endpoint results. Two different restriction enzymes resulting in two digest patterns are used in this section for analysing recombination reactions - they provide unique insights into the distribution of recombination products (Fig. 4.8). These are clearly annotated where necessary.

The 6 TALERs showed significant recombination activity on IVTS22 and IVTS24 with observable resolution products (Fig. 4.9). In contrast to Holt (2014), resolution products were observed with TALER6X on IVTS22 and IVTS24. Across the board, IVTS22 seems to be the most favoured substrate. Activity on the substrates reduced with increasing spacer length and no definitive resolution products were observed on IVTS28. Linker length does not seem to have a significant impact on $\Delta 148$ truncated TALER activity. The resolvase cut-off point also seems to have minimal impact with TALER6X (R144- $\Delta 148$) behaving quite like TALER6 (R148- $\Delta 148$) except for a minor increase in resolution products with IVTS24. The protein concentrations used here were estimated using SDS-PAGE. In subsequent experiments, protein concentrations were estimated using the Bradford assay to ensure fair and standard comparison.

Since increasing the length of the flexible linker did not significantly alter recombination activity, it was predicted that a rigid linker might structurally and functionally separate the TALE DBD from the resolvase catalytic domain by increasing the stiffness and increase the amount of resolution products generated. A 12-aa rigid linker (AEAAAKEAAKA'TS') $\Delta 148$ TALER variant with sequence based on the 'A(EAAK)_nA' alpha-helix forming rigid linker design (Chen *et al.*, 2013) was expressed, purified and tested. The TALER was not as active TALER6 or TALER12 (data not shown).

Three more TALER truncation variants were expressed and purified - $\Delta 221$, $\Delta 153$ and $\Delta 119$, (Fig. 4.6 and Fig. 4.7). $\Delta 221$ and $\Delta 153$ were tested to analyse if truncating the N-terminal TALE DBD further will improve activity and $\Delta 119$

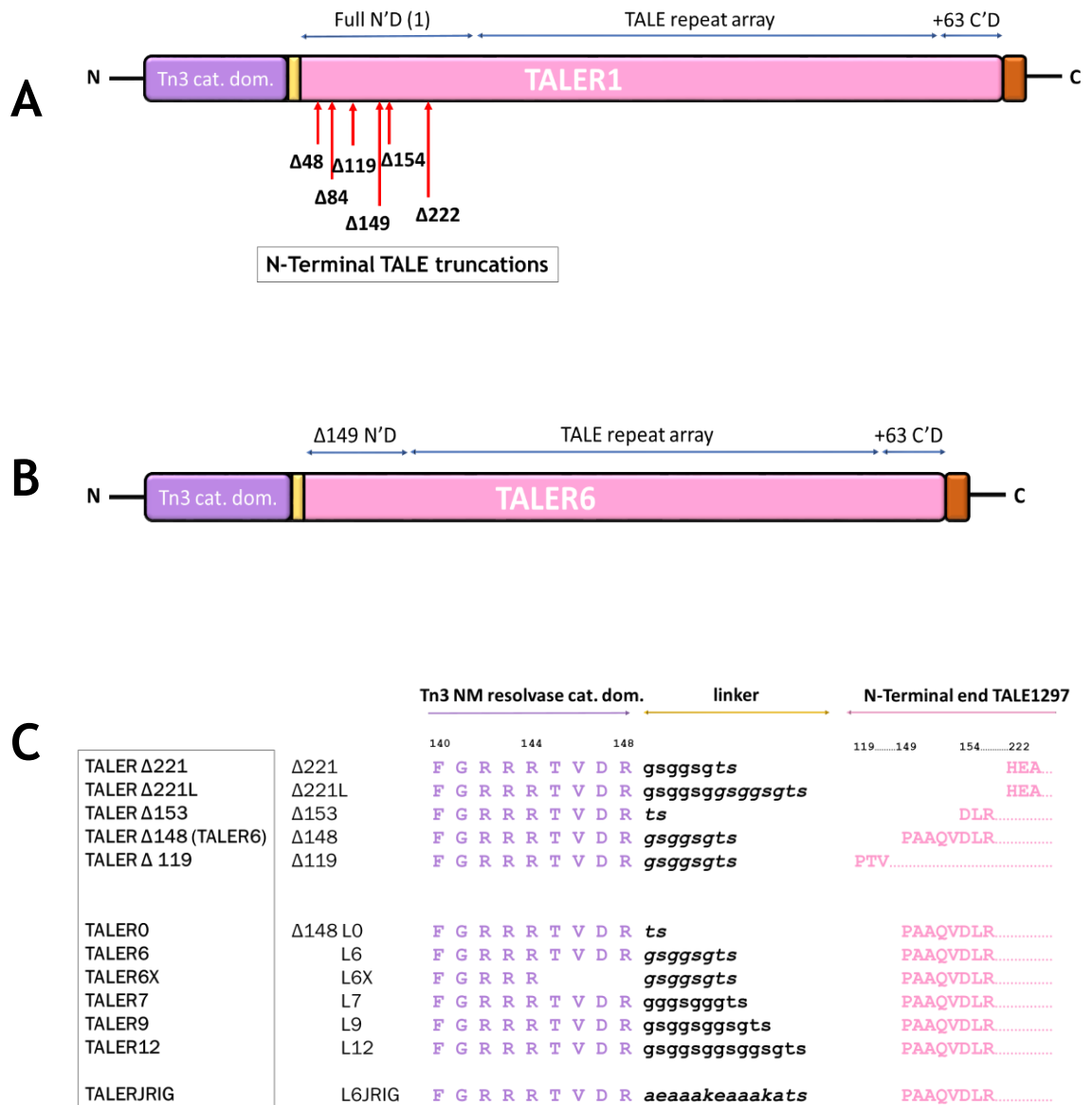
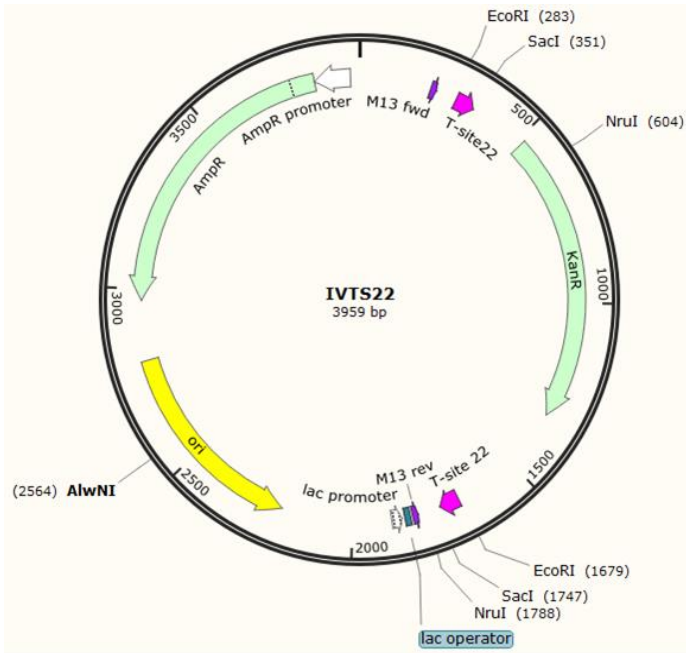


Figure 4.6: Truncation and linker length variants used in this work. A. TALER1 is the full length TALE1297 fused to the C-terminal end of Tn3 NM resolvase (cut-off at residue 144) by a 6-aa flexible linker. B. The basic structure of TALER6 ($\Delta 148$ with 6-aa GS linker) is shown. C. Four N-terminal truncations of TALER1 are explored in this work- $\Delta 221$, $\Delta 153$, $\Delta 148$ and $\Delta 119$. An additional $\Delta 221$ TALER with a 12aa linker ($\Delta 221L$) was tested. Seven variants of the $\Delta 148$ N-terminal truncation were generated. These TALERs are named according to their linker length TALER6 has a 6-aa GS linker, TALER 12 has a 12-aa GS linker and TALERO has no GS linker. The short 'TS' linker is ignored in this numbering. TALERJRIG has a 12-aa rigid linker. Asides form TALER6X, all TALER variants used in this section were cloned in the course of this research work by carefully designing annealed oligonucleotides to introduce the required changes.

MAIFGYARVSTSQQSLDIQIRALKDAGVKANRIFTDKASGSSTDREGLDLLRMKVKEGDVILVKKLDRLGRDTADMIQLIKEFDAQGVAVRFIDDGISTDS
 YIGLMVVTILSAVAQAERRRILERTNEGRQEAKLKGIKFGRRTVDRGSGSGSTS^{144 148 2}DP IRSRTPSPARELLPGPQPDSVQPTADRGGAPPAGGPLDGLPARR
 49 85 120
 TMSRTRLPSPPAPSPAFSAGSFSDLLRQFDPSLLDTSLLDSMPAVGTPHTAAAPAECDEVQSGLRAADDPPP^{149 154}TVRVAVTAARPPRAKPAPRRRAAQPSDAS
 222
 EAAQVDLRTLGYSQQQQEKIKPKVRSSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELR
 GPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLNLTDPQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNNGGKQALETVQRLLPV
 LCQAHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPAQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLP
 VLCQDHGLTPEQVVAIANNNGGKQALETVQRLLPVLCQAHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPAQVVAIASNNGGKQALETVQRLLP
 PVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPDQVVAIASHDGGKQALETVQRL
 LPVLCQAHGLTPAQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNIGGKQALETVQR
 LLPVLCQAHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPAQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNNGGGRPALESIV
 AQLSRPDPALAAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAHHHHHH*

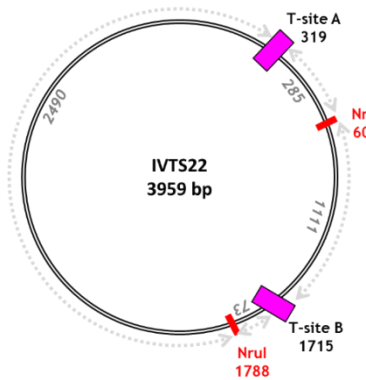
Figure 4.7: Sequence of TALER1 with truncation points annotated. The sequence of Tn3 NM resolvase is linked to the N-terminal region of the TALE domain (featuring N-terminal region sequences from PthXo1 up to TALE residue 152, and CRD with +63 C-terminal sequences from TALE1297) using a - aa GS linker with a SpeI restriction site (TS). Tn3 NM resolvase residues are highlighted in grey and numbered in italics. TALE DBD residues are numbered in bold; the letters highlighted in red show the starting residues of the TALE DBD in the truncated variants used by Holt (2014) and in this work. The four cryptic repeats of the TALE NTR are coloured based on the crystal structures shown in Fig. 30 and are based on Gao *et al.*, (2012). TALE RVD repeats are highlighted in different shades of blue and the C-terminal region (+63) is in green.

A



B

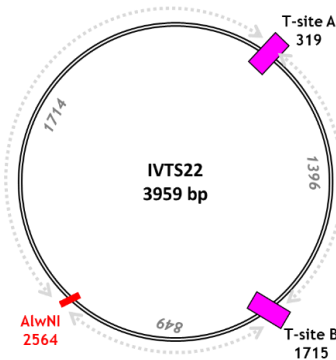
NruI digest pattern



UR - Unrecombined:	1184, 2775
RP - Resolution product:	1396, 2563
IP - Inversion product:	358, 3601
CP - Cleavage products	
Double-site cleavage:	73, 285, 1111, 2490
Single-site cleavage at T-site A:	285, 1184, 2490
Single-site cleavage at T-site B:	73, 1111, 2775

C

AlwNI digest pattern



UR - Unrecombined:	3959
RP - Resolution product:	1396, 2563
IP - Inversion product:	3959
CP - Cleavage products:	
Double-site cleavage:	849, 1396, 1714
Single-site cleavage at T-site A:	1714, 2245
Single-site cleavage at T-site B:	849, 3110

Figure 4.8: Restriction patterns for analysis of *in vitro* reactions. A. The plasmid map of IVTS22 shows the two T-sites flanking a *kanR* gene in direct repeat. B. The NruI digest pattern allows for the visualization of the inversion product (3601 bp) and the second resolution product (1396) distinctly from other bands. However, the first resolution product band is obscured by closely running cleavage products. C. The AlwNI digest pattern allows for the visualization of this first resolution product band although the second resolution product and the inversion products can be obscured here.

(a variant previously analysed by Holt, 2014) to retest its activity (Fig. 4.10). Again, $\Delta 148$ (TALER6) seemed to be the optimal truncation point for TALE1297 with the 6-aa truncation variant ($\Delta 153$) losing most excision activity. Similar to what is observed with linker lengths, there does not seem to be a correlation between the extent of the TALER (N-terminal TALE) truncation and target site spacer length. The longer TALER variant ($\Delta 119$) is just as inactive on IVTS28 as on IVTS22.

IVTS22 was selected as the main substrate for most of the subsequent work here as it aligns perfectly with the CR_TATA_target site allowing for the positioning of the T-sites with the canonical 5' T at position 0 for HIV-TALE DBD protein design (Section 5.2.2). Analysis of all the TALER variants on IVTS22 using the AlwNI restriction digest pattern shows the recombination product distributions (Fig. 4.11). The results here are synonymous to what has been previously described above, with $\Delta 148$ TALERs yielding the best activity. TALER6 was also chosen as the optimal variant as it had the shortest linker length without the complexities of TALER0 insolubility during purification. The distribution of its recombination products on IVTS22, IVTS24, IVTS26 and IVTS28 is shown in Figure 4.12 and is as previously described, with IVTS22 and IVTS24 as optimally-spaced substrates.

Catalytic domain variants of TALER6 were generated. This was to confirm if reducing the activating mutations in the Tn3 NM variants would improve the recombination activity of TALER6. Three previously-characterised Tn3 resolvase variants were swapped with the NM resolvase catalytic domain present in TALER6 (NM R148-6aalinker- $\Delta 148$ TALE). The Tn3 catalytic domains used were the wild-type resolvase catalytic domain (T6-WTR), the 'SY' activated mutant variant with mutations G101S and D102Y (T6-SY), and the 'M' activated mutant variant with mutations G101S, D102Y, M103I, Q105L (T6-M) (Olorunniji *et al.*, 2008). Recombination analysis shows that the original NM variant (TALER6) is still the most active (Fig. 4.13). In contrast to previously described results, the 'M' catalytic domain shows no activity here. This could be because of the reduced solubility of T6-M as it consistently precipitated out of its storage solution (here, 1M NaCl and 50% glycerol).

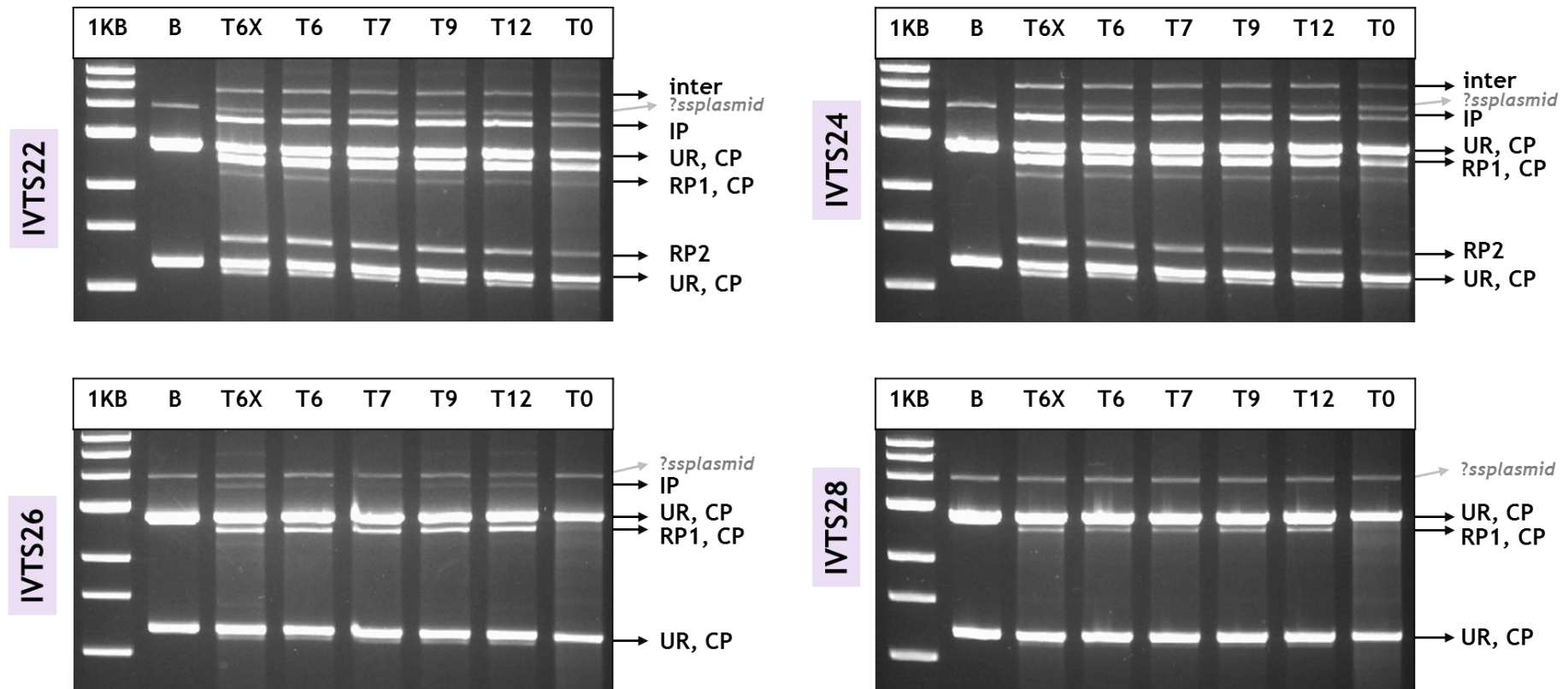


Figure 4.9: *In vitro* reaction of TALER linker length variants with $\Delta 148$ N-terminal truncation on IVTS22, IVTS24, IVTS26 and IVTS28 (NruI-digested). The smaller resolution product, RP2, is definitive of recombination activity as described in Fig. 8. Resolution products are not observed on IVTS26 and IVTS28 although some cleavage products can be seen. Protein concentration added to the reactions: 1/8 protein dilution of TALER6X (about 4 μM) and its equivalent as estimated on denaturing SDS-PAGE gel (Section 2.13). DNA concentration in the reactions: 50 $\mu\text{g}/\text{ml}$. The band marked “*?ssplasmid” is unidentified and could be a single-stranded DNA plasmid.

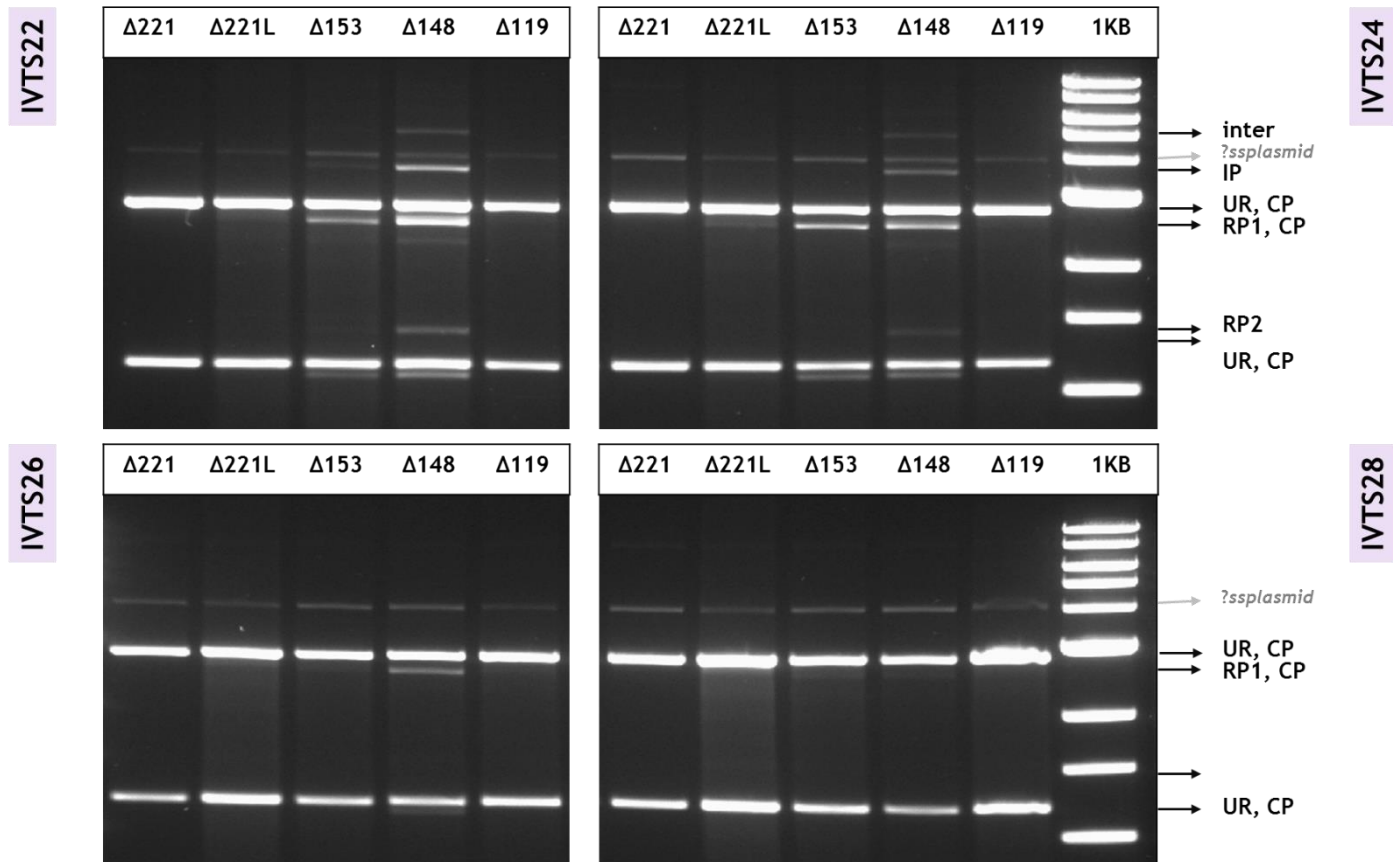


Figure 4.10: *In vitro* reaction of N-terminal truncation TALER variants on IVTS22, IVTS24, IVTS26 and IVTS28 (NruI-digested). The smaller resolution product, RP2, is definitive of recombination activity as described in Fig. 8. Resolution products are only observed on IVTS22 and IVTS24 with the $\Delta 148$ truncation variant. Cleavage products can also be seen with $\Delta 153$ on IVTS22 and IVTS24. The other truncation variants do not seem to demonstrate any recombination activity. *Protein concentration in the reactions: 400 nM; DNA concentration in the reactions: 20 $\mu\text{g/ml}$*

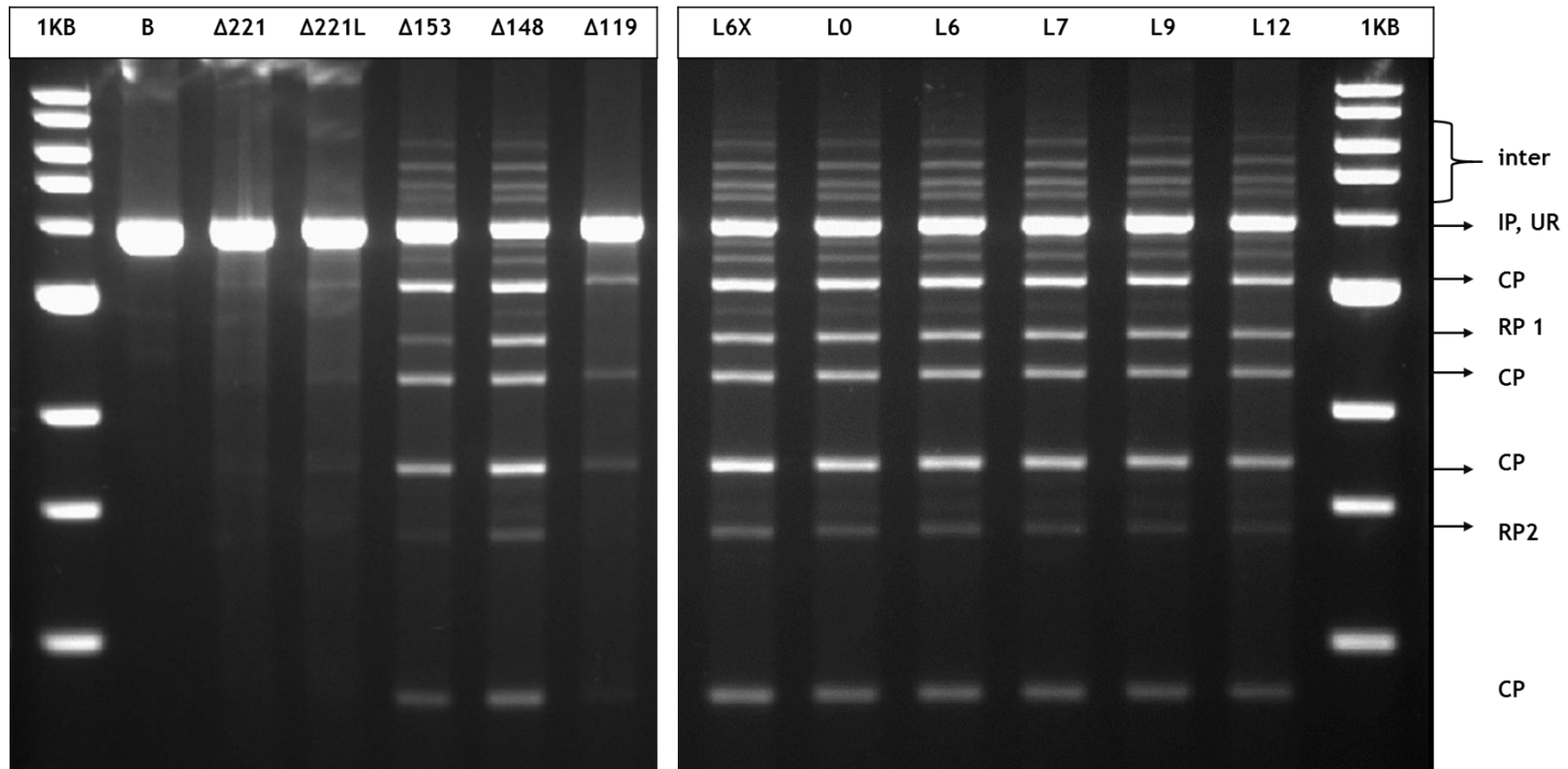


Figure 4.11: Recombination product distribution from *in vitro* activities of N-terminal truncation and linker length TALER variants on IVTS22 (AlwNI-digested). The bigger resolution product, RP1, is definitive of recombination activity as described in Fig. 8. The $\Delta 221$ variants do not demonstrate significant recombination activity. Resolution products are observed with the $\Delta 148$ truncation variant and changes in linker length do not yield significant differences in activity. Some resolution products can be observed with the $\Delta 153$ TALER although the products here seem to be predominantly from cleavage. Some cleavage products can also be seen with the $\Delta 119$ TALER variant.

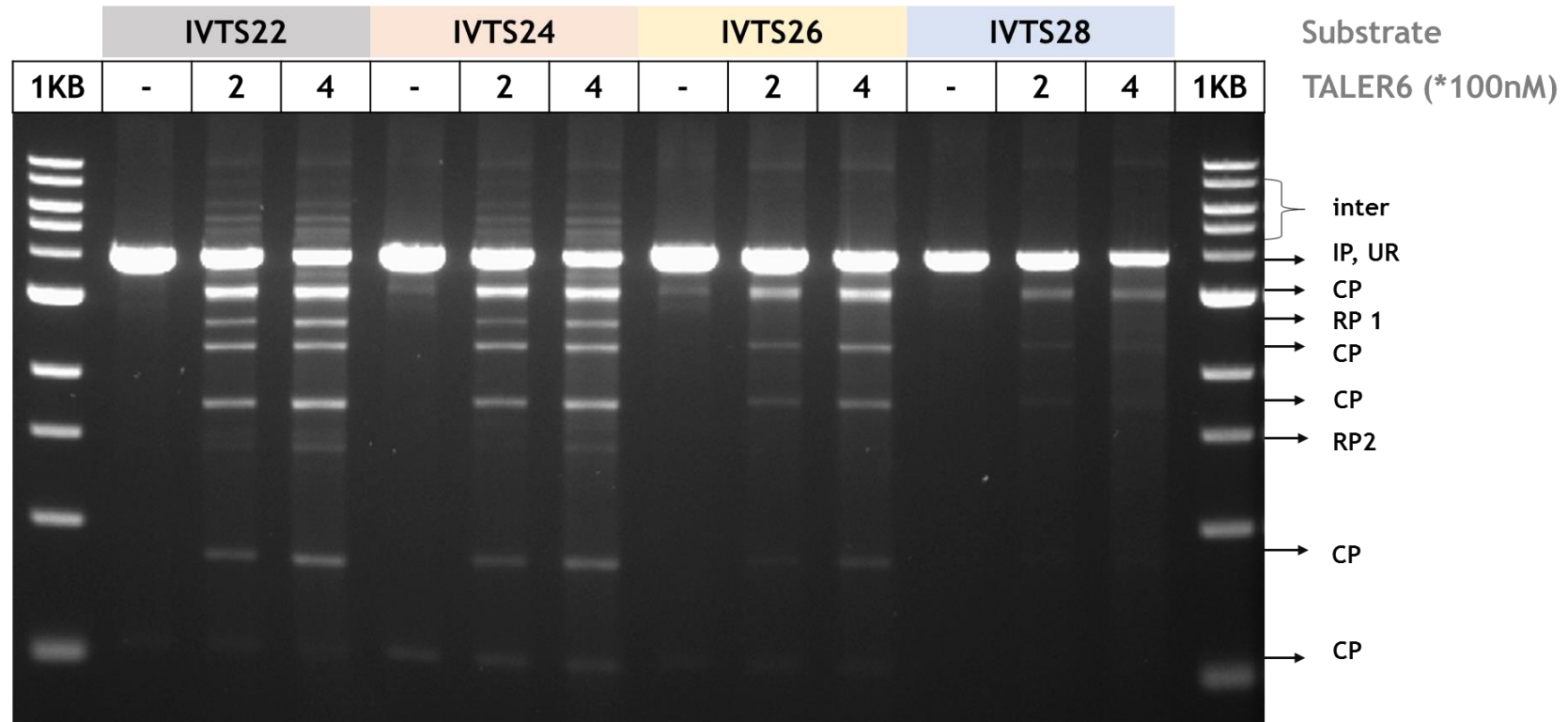


Figure 4.12: Recombination product distribution from *in vitro* activities of TALER6 on IVTS22, IVTS24, IVTS26 and IVTS28 (AlwNI-digested). The bigger resolution product, RP1, is definitive of recombination activity as described in Fig. 8. Resolution products are observed with IVTS22 and IVTS24 although a minor preference seems to be for IVTS22. No resolution products can be observed with the IVTS26 and IVTS28 although some cleavage products can be observed.

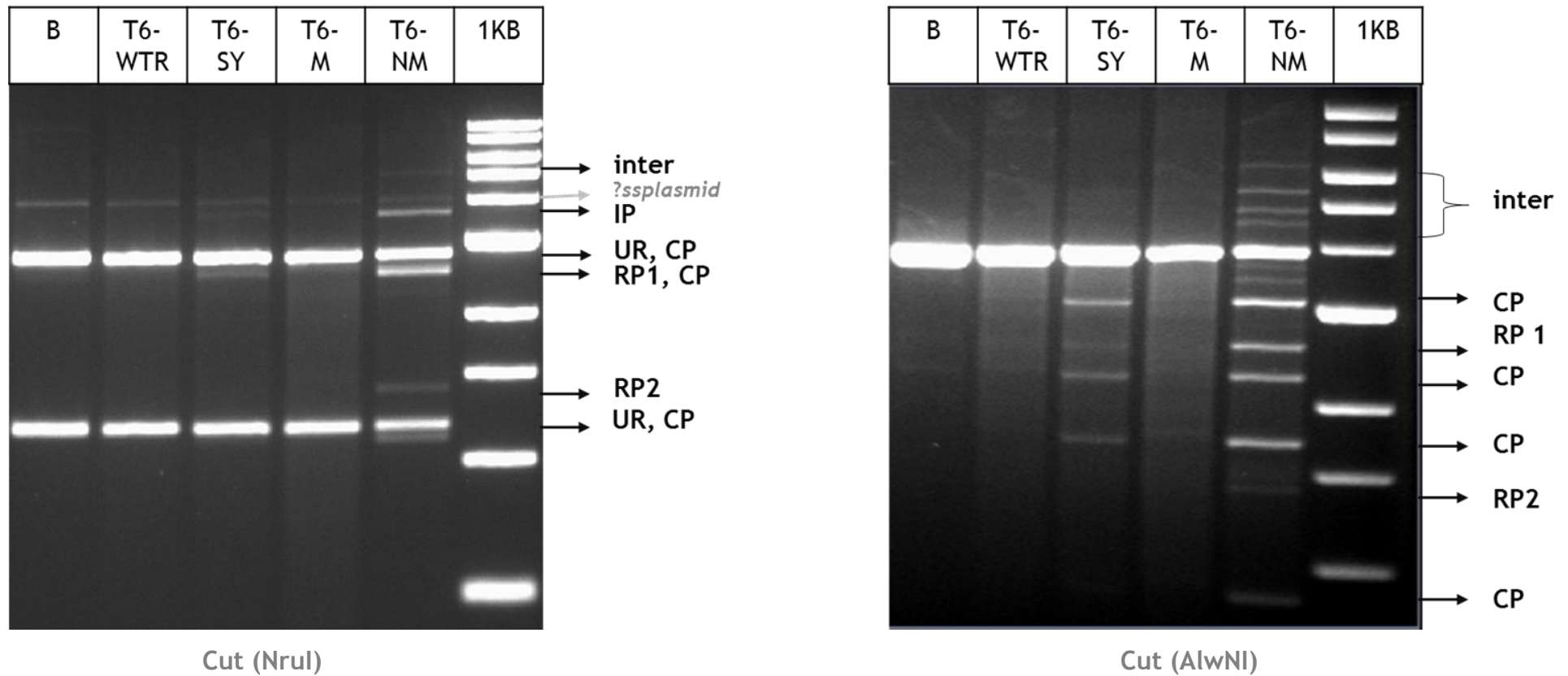


Figure 4.13: *In vitro* activities of TALER6 catalytic domain variants on IVTS22. No recombination activity is observed with the T6-WTR as expected. The standard TALER6, T6-NM (R2A, E56K, G101S, D102Y, M103I, Q105L) shows the most activity, T6-M (G101S, D102Y, M103I, Q105L) is inactive and only cleavage activity can be observed with T6-SY (G101S, D102Y). Resolution products are observed with IVTS22 and IVTS24 although a minor preference seems to be for IVTS22.

4.2.3 *In vitro* characterization of TALER activity

Several approaches were taken to improve the activity of TALER6 on IVTS22. This included the characterization of TALER6 activity under varying magnesium and salt concentrations. These did not seem to improve the recombination product distribution significantly (Fig. 4.14). The effect of ethylene glycol and EDTA on TALER recombination activity was also considered (Fig. 4.15). TALER behaviours under varying pH conditions are as described for activated mutants of Tn3 resolvase, showing significant activity at lower pH but not much disparity in the formation of resolution products from pH 6.0 to pH 10.0 (data not shown). The standard reaction conditions described in Section 2.16 were sufficient for maximal activity.

Protein-DNA concentration analysis shows that the distribution of recombination products was similar across varying DNA concentrations (from 10 µg/ml to 50 µg/ml) and protein dilutions (400, 800 and 1600 nM) (Fig. 4.16). Further analysis, using a shorter reaction time and diluting the reaction products to the same concentration before agarose gel analysis showed subtle differences in activity. 2 nM TALER6 on 20 µg/ml of IVTS22 provided a good level of resolution products with fewer intermolecular recombination products than when higher concentrations of TALER and DNA are used (data not shown).

Schreiber *et al.*, (2015) indicated that DNA-free AvrBs3, a TALE protein, formed multimeric complexes *in vitro* that were detrimental to DNA binding by interacting with itself through disulphide bonds. Two cysteines present in the C-terminal domain of AvrBs3 are implicated in this intermolecular interaction. Schreiber and colleagues showed that pre-treating the protein with dithiothreitol (DTT) dissociated these complexes. Since most TALE proteins have similar architectures, TALE 1297 has a similar sequence to AvrBs3. However, in the C-terminal truncation variant used in this work (+63), only one of those cysteines is retained. Disulphide-bridge formation could reduce the availability of TALERs for recombination or drive intermolecular recombination by stabilizing illicit complexes. To analyse the implication of this potential multimerization, TALER6

was treated with 0 mM, 5 mM, and 10 mM DTT overnight at 4 °C and then recombination reactions on IVTS22 were carried out as before. DTT pre-treatment did not seem to influence TALER activity and this could imply that intermolecular recombination is not driven by disulphide bridge formation (Fig. 4.17).

Analysing the reaction times for TALER6 on IVTS22 shows that cleavage under ethylene glycol conditions seems to proceed quite fast, with significant substrate depletion observed from 2 minutes of reaction and double-site cleavage products appearing from 5 minutes (Fig. 4.18). Resolution products seem to be observable under these ethylene glycol conditions, as evidenced by what appears to be a free circle running faster than the smaller cleavage product on the gel. Recombination (under standard conditions) progresses steadily from 1 minute and intermolecular recombination products begin to appear at 15 minutes (Fig. 4.19a). This coincides with the appearance of resolution and inversion products which can be observed after the reactions are digested (Fig. 4.19b). Single-cut DNA (linear) appears to not be processed further as it can be observed to stay steady across increasing reaction times; with most of the substrate being converted from supercoiled plasmid DNA into recombination products. This might infer that substrate supercoiling drives the recombination reaction forward as it does with wild-type Tn3 resolvase (Stark *et al.*, 1994)

To evaluate the effect of substrate supercoiling on recombination activity, the activity of TALER6 on supercoiled and linear substrates of IVTS22 was observed. The linear substrate was obtained by linearizing IVTS22 with AlwNI (Fig. 4.20). Resolution products from supercoiled substrates were notably more than that with linear substrates. Thus, DNA supercoiling seems to be important for recombination.

Product distributions from recombination reactions with TALER6 have shown significant levels of intermolecular recombination and an accumulation of cleavage products which could result from such intermolecular recombination. To probe intermolecular recombination, varying protein concentrations of TALER6

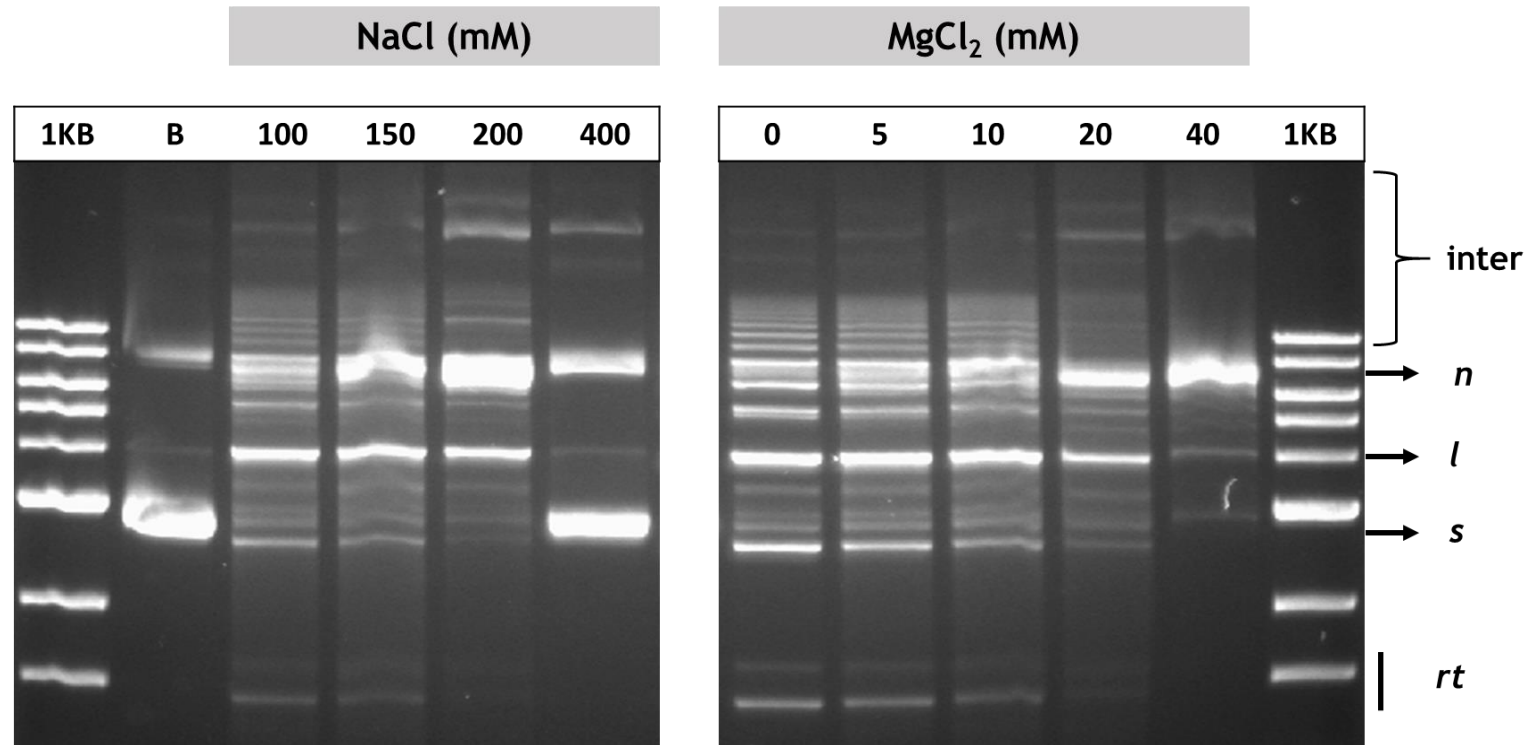


Figure 4.14: Effect of increasing concentrations of NaCl and MgCl₂ on the *in vitro* activity of TALER6 on IVTS22. These reactions were carried out under standard recombination reaction conditions, with only the NaCl or MgCl₂ altered as indicated (Section 2.16). Increasing the concentrations of NaCl from the standard reaction concentration (100 mM) leads to an increase in nicking activity of the protein. At 400 mM NaCl concentration, TALER6 is no longer active. The standard reaction concentration of MgCl₂ is 10 mM. Reducing the concentration leads to an accumulation of cleavage products while increasing the concentration leads to increased nicking activity and loss of substrate plasmid. The abbreviations are as follows: “n” (nicked substrate), “l” (linear substrate), “s” (supercoiled substrate), “rt” (recombination topoisomers), “inter” (intermolecular recombination products).

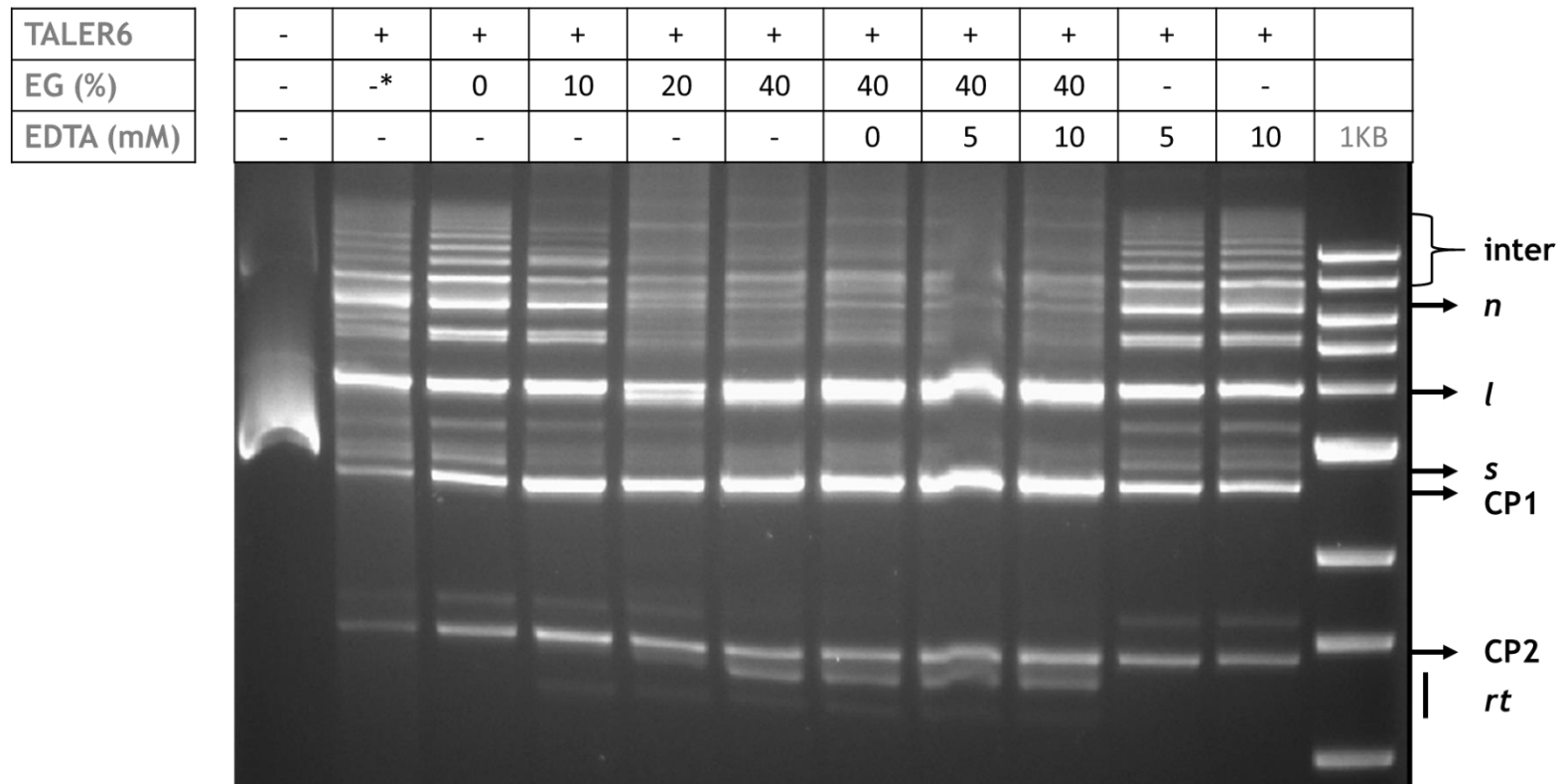


Figure 4.15: Effect of ethylene glycol (EG) and EDTA on the *in vitro* activity of TALER6 on IVTS22. These reactions were carried out under standard recombination reaction conditions without $MgCl_2$ (except in lane 2*, where $MgCl_2$ was added to 10 mM) and with ethylene glycol and/or EDTA added as indicated (Section 2.16). The addition of increasing concentrations of EG leads to an increase in the appearance of cleavage products. At 40% EG, there is the appearance of a faster running band which could be the free circle of resolution product 2. EDTA does not seem to influence the reaction significantly. The abbreviations are as follows: “n” (nicked substrate), “l” (linear substrate), “s” (supercoiled substrate), “CP” (double-site cleavage products), “rt” (recombination topoisomers), “inter” (intermolecular recombination products).

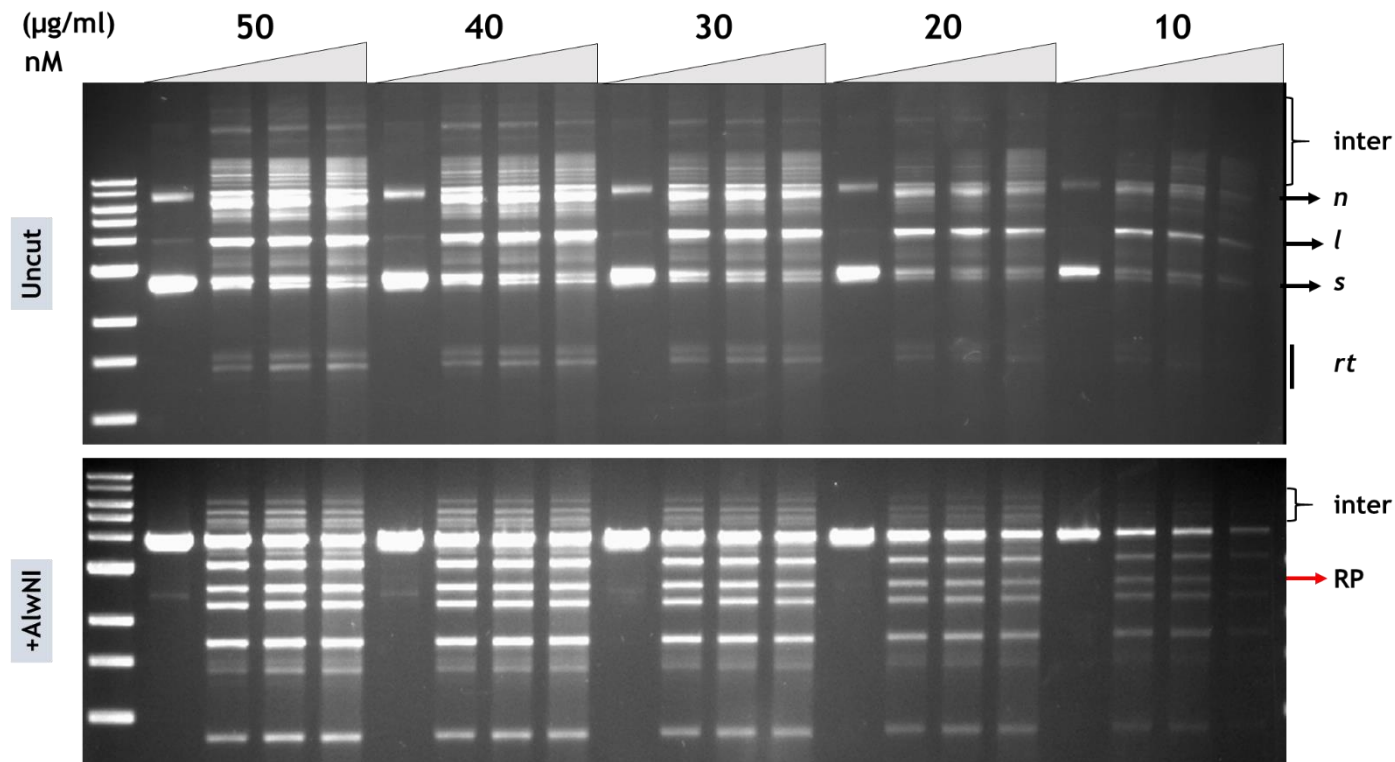


Figure 4.16: Effect of protein-DNA concentration on *in vitro* activity of TALER6X on IVTS22. Increasing the concentrations of the TALER6X does not seem to shift recombination product distribution significantly. The uncut reaction is shown at the top and the AlwNI-digested reactions at the bottom. Five substrate DNA concentrations were tested (10, 20, 30, 40 and 50 µg/ml), each with protein concentrations (0, 400, 800 and 1600 nM). The abbreviations are as described in Figure 4.15. RP2 which is definitive of recombination activity as described in Fig. 8 is indicated as RP. Product distribution is fairly similar across the protein-DNA concentration ranges tested.

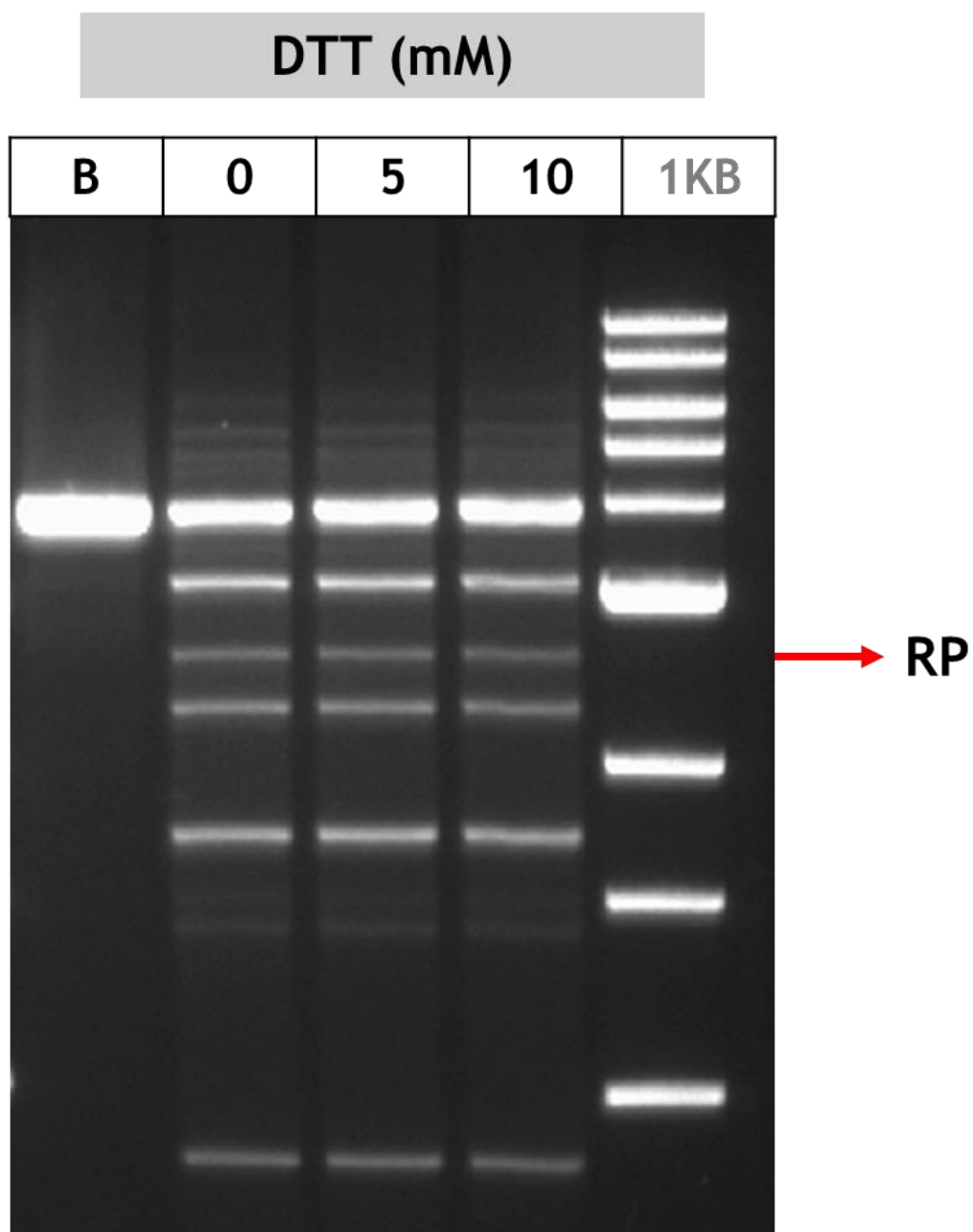


Figure 4.17: Effect of DTT pre-treatment on the *in vitro* activity of TALER6 on IVTS22. These reactions were carried out under standard recombination reaction conditions (Section 2.16). All proteins used were incubated with the indicated amount of DTT overnight at 4 °C before the recombination reaction. No increase in resolution products or other reaction products is observed.

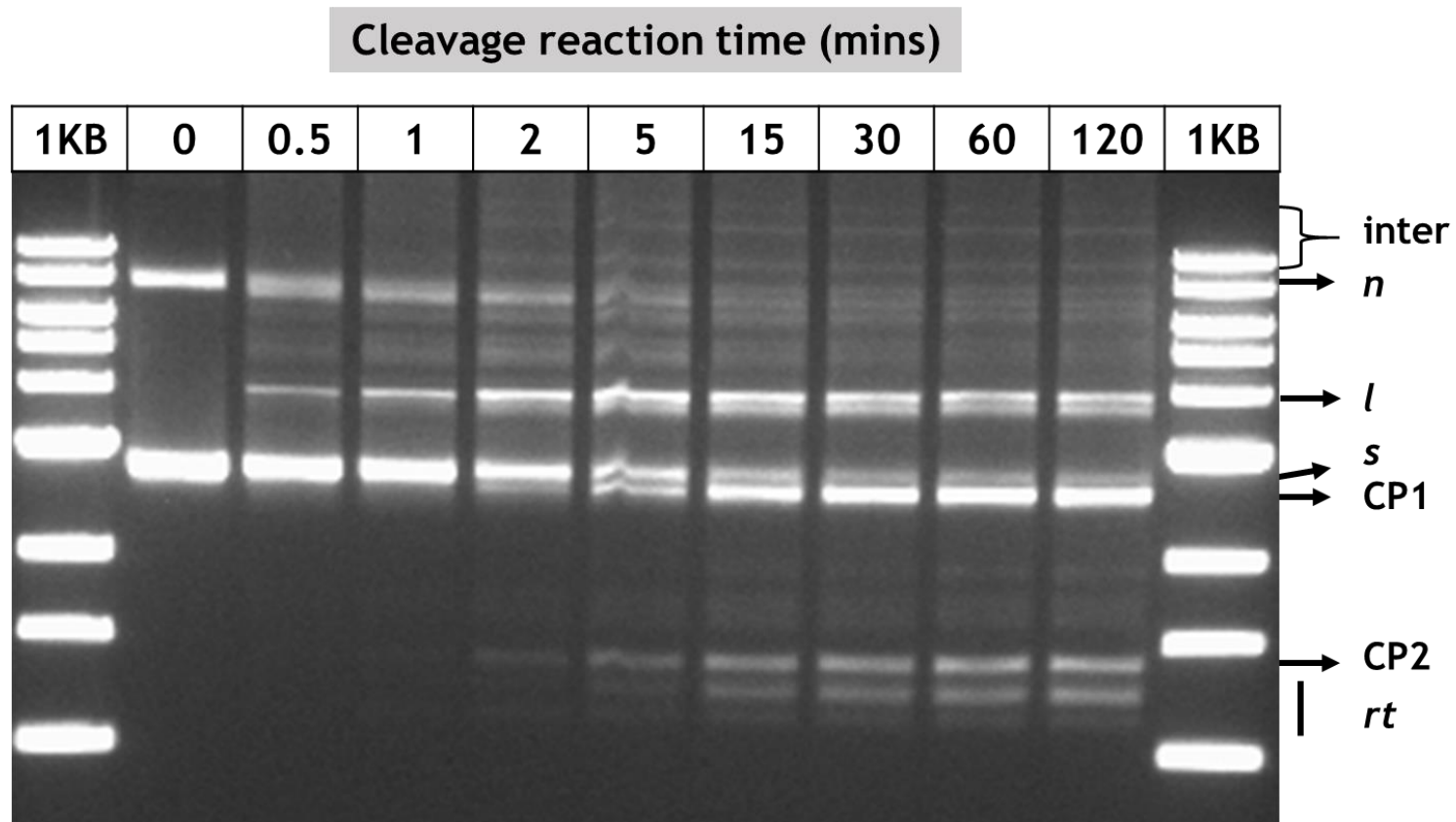


Figure 4.18: Time course of *in vitro* cleavage activity of TALER6 on IVTS22. These reactions were carried out under standard cleavage reaction conditions with 40% ethylene glycol (Section 2.16). There is an increase in double-site cleavage products with time. Resolution topoisomers (rt) also appear signifying resolution under cleavage conditions.

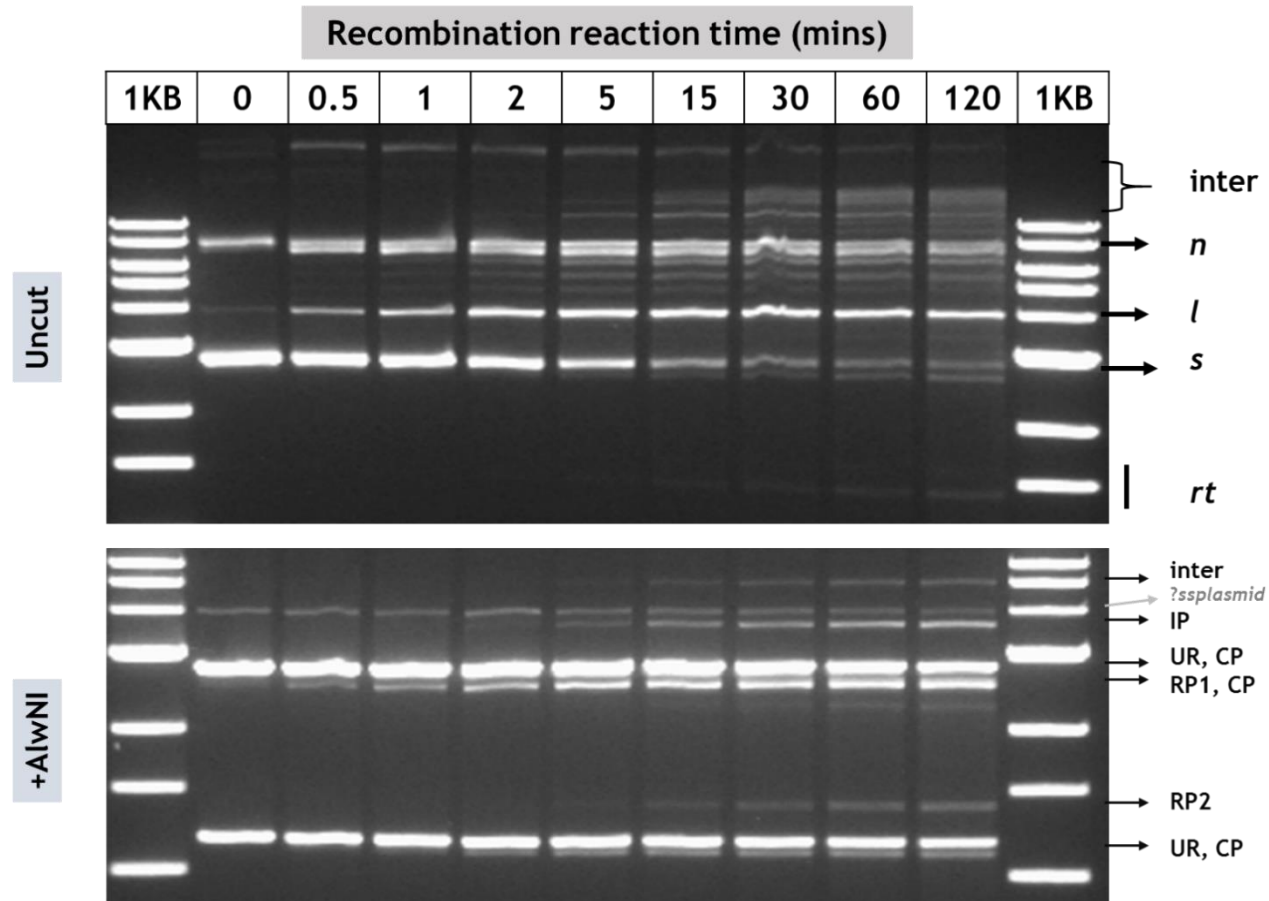


Figure 4.19: Time course of *in vitro* recombination activity of TALER6 on IVTS22. These reactions were carried out under standard recombination reaction conditions (Section 2.16). There is an increase in resolution products and intermolecular recombination products with time.

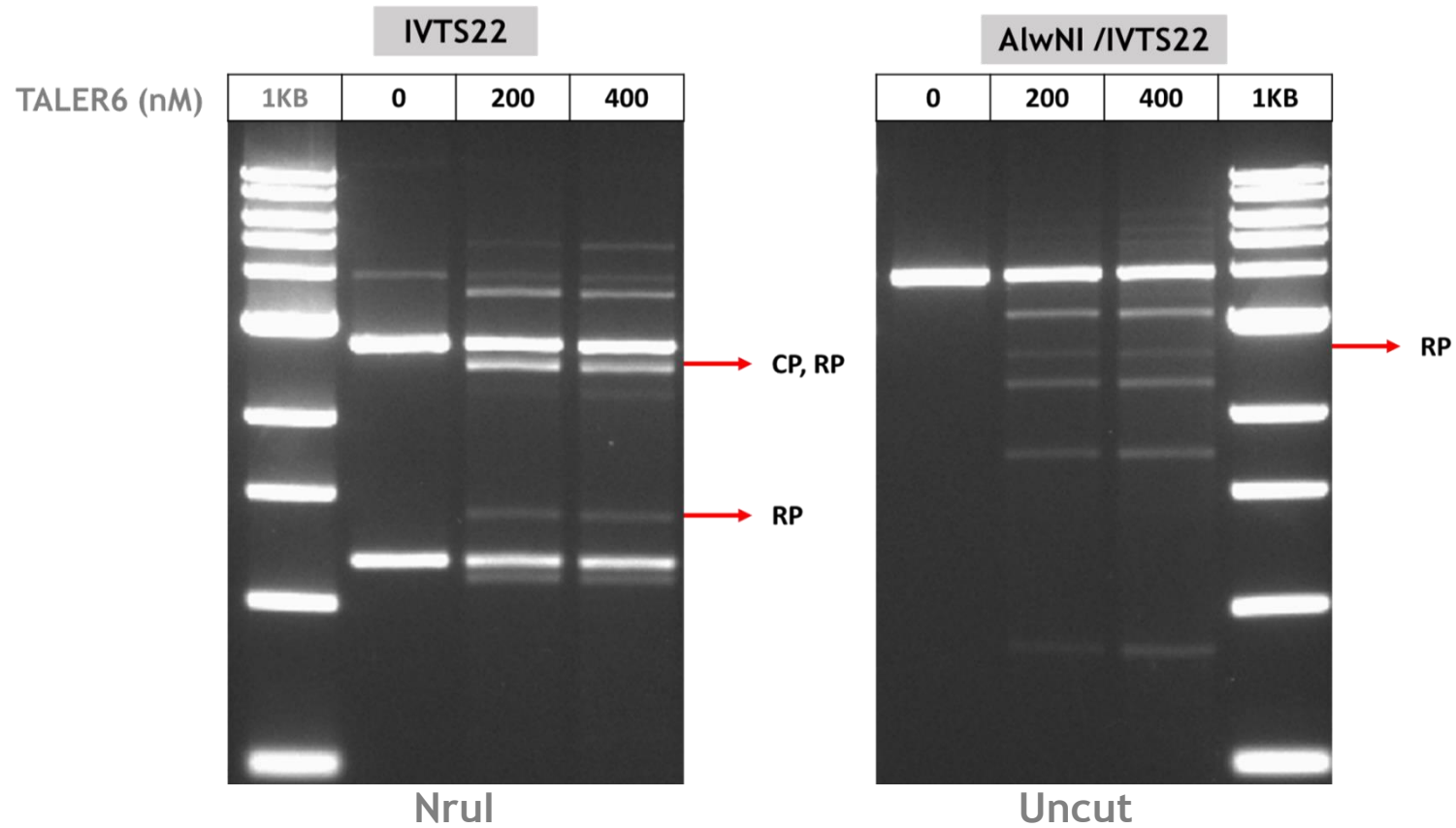


Figure 4.20: Effect of supercoiling on the *in vitro* activity of TALER6 on IVTS22. The activity of TALER6 on supercoiled substrate plasmid (IVTS22) (digested with NruI) and linear DNA (AlwNI/IVTS22) generated from the digestion of IVTS22 by AlwNI, is compared.

were analysed on single-site substrate plasmids. Recombination of these substrates should lead to the integration of the two plasmids. The activity of TALER6 on T1S22, T1S24, T1S26 and T1S28 is akin to what is seen with double-site substrates (Fig. 4.21). The digest using AlwNI here shows a significant accumulation of cleavage products along with the integration products where present. The products on T1S26 are predominantly cleavage products. TALER6 catalyses intermolecular recombination on T1S22 and T1S24. Increasing levels of integration products were observed as DNA concentration increases (data not shown).

It was predicted that temperature changes might alter the product distribution from recombination. A lower reaction temperature might reduce random collision of DNA molecules and allow intramolecular recombination to proceed faster than intermolecular recombination, while a higher reaction temperature might improve the rate of product conversion. The activity of TALER6 on IVTS22 was tested at temperatures ranging from 25 °C to 60 °C. TALER6 showed recombination activity at all temperatures tested although substrate conversion reduced with increase in temperature (Fig. 4.22). Product distribution was not significantly altered; however, 37 °C (standard reaction temperature) yielded the maximum amount of resolution products.

Comparing the activity of TALER6 side-by-side with Tn3 NM resolvase and Tn3 NM ZFR showed that although the substrate plasmids are substantially depleted across all three proteins, there is an accumulation of single-cleaved linear products with TALER6 that is not observed with NM resolvase and is seen only minimally with NM ZFR (Fig 4.23 and Fig. 4.24). This explains the accumulation of cleavage products observed after restriction digests. To assess the capacity of TALER6 to be complemented by Tn3 NM resolvase or NM ZFR by forming heterodimeric synaptic complexes, two substrate plasmids, carrying one T-site as site A and either Tn3 *res* site I or Tn3 Z-site as site B, were designed. A complementation assay showed no cooperativity between TALER6 and either Tn3 NM resolvase or Tn3 NM ZFR (data not shown).

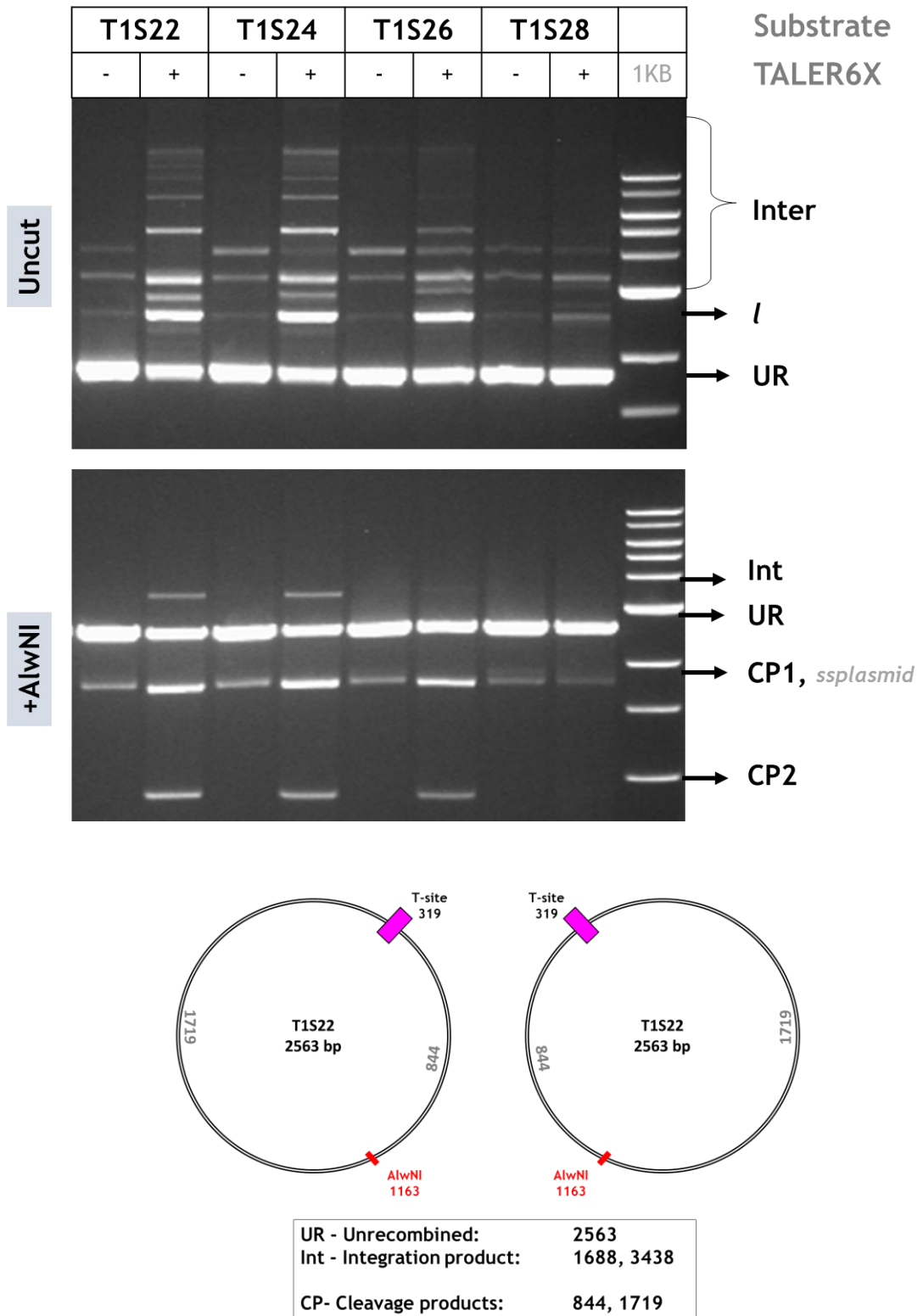


Figure 4.21: Analysis of integration activity of TALER6 on single-site substrate, T1S22. NruI digest shows the accumulation of linear cleavage products as by-products of integration by TALER6. Integration products are annotated.

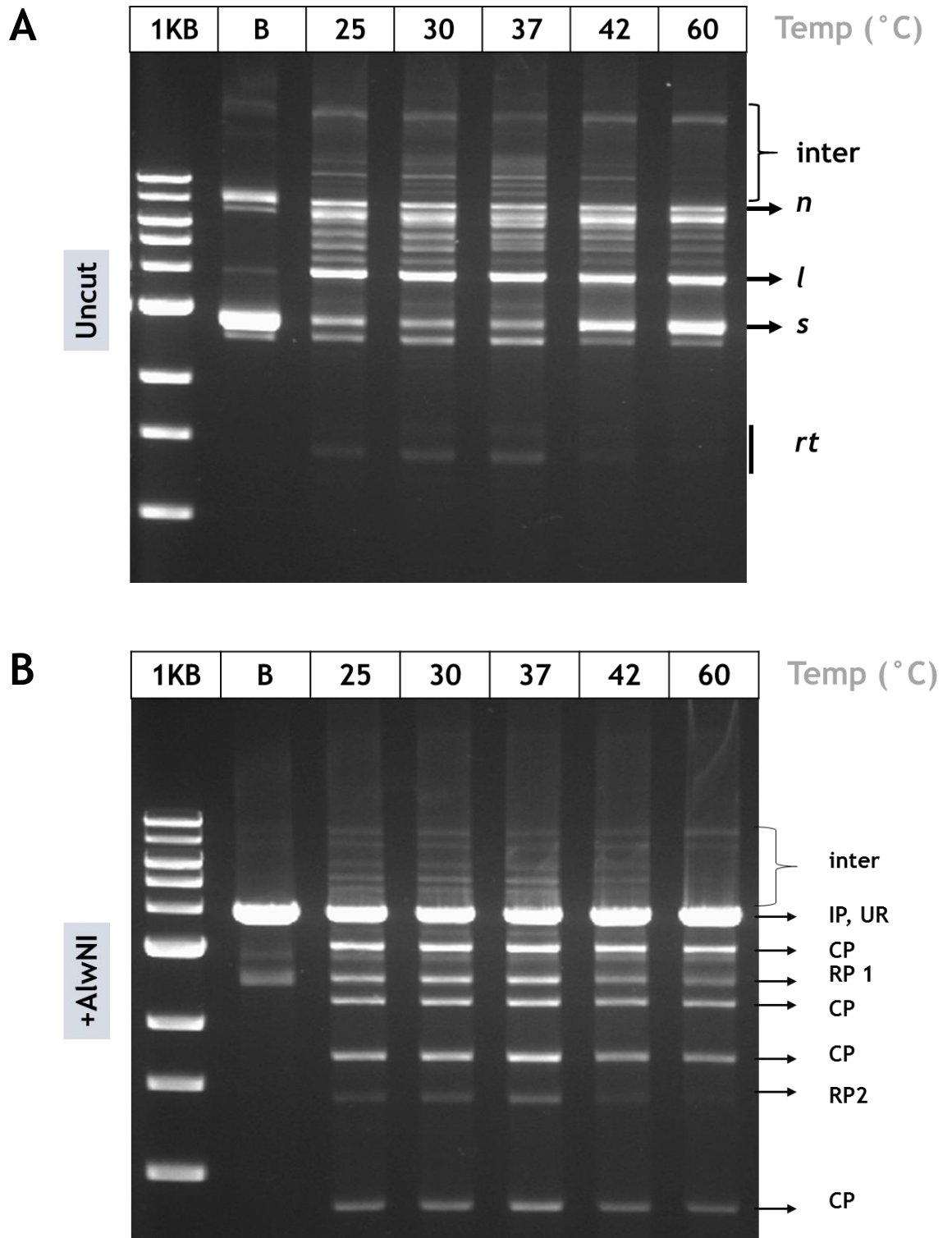


Figure 4.22: Effect of temperature on the *in vitro* activity of TALER6 on IVTS22. A. Uncut. B. AlwNI-digested. TALER6 is active across all temperatures observed although substrate conversion seems to be significantly reduced from 42 °C. 37 °C seems to yield the highest amount of recombination products. Product distribution does not seem to be altered by temperature changes. These reactions were carried out under standard recombination reaction conditions (Section 2.16).

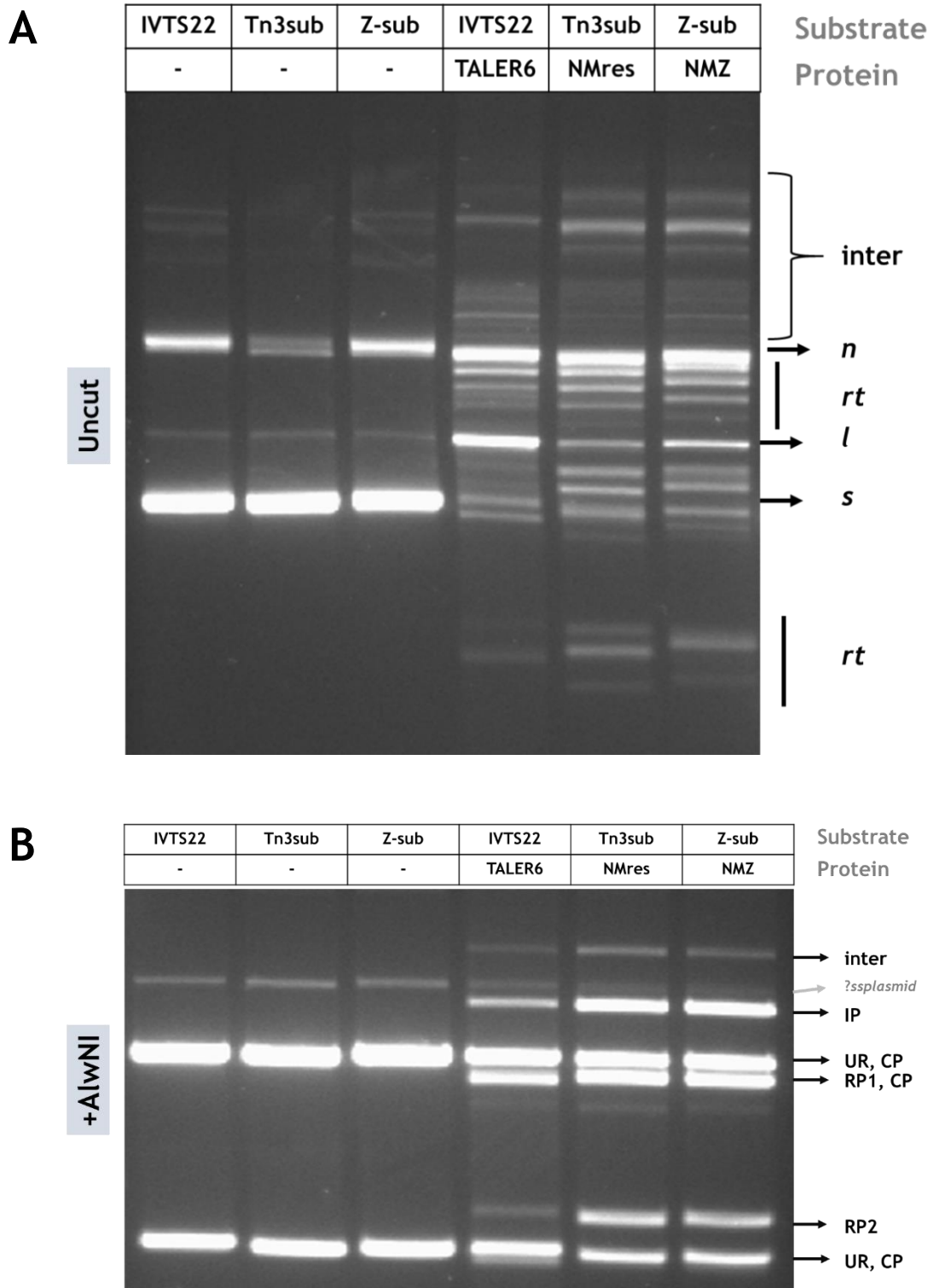


Figure 4.23: Comparison of TALER6 activity on IVTS22 with that of Tn3 NM resolvase and Tn3 NM ZFR on their own substrates. A. Uncut. B. AlwNI-digested. The substrate plasmids of Tn3 NM resolvase and Tn3 NM ZFR are Tn3sub (pMP78) and Z-sub (pDWIVS6). Tn3 NM resolvase and Tn3 NMZ yield significantly more resolution products than TALER6. Although they all show substrate depletion to the same extent, the accumulation of single-site cleavage products is characteristic of TALER6 activity. These reactions were carried out under standard recombination reaction conditions (Section 2.16).

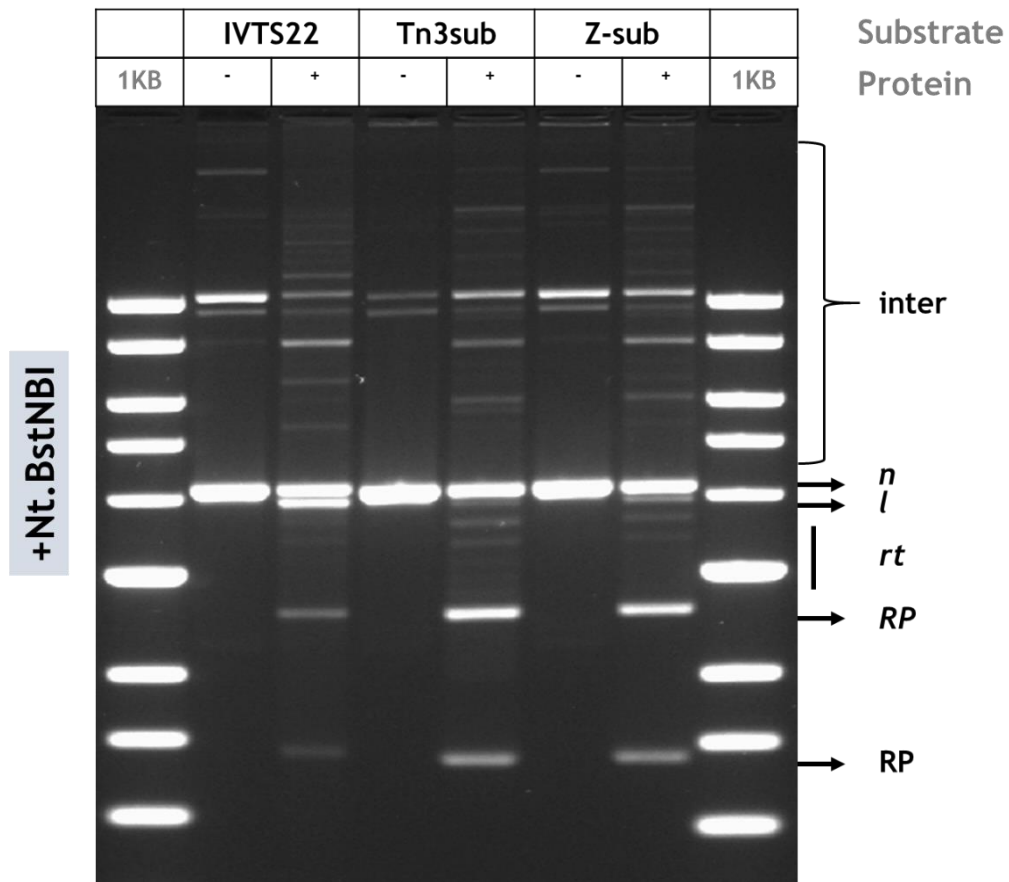


Figure 4.24: Comparison of TALER6 activity on IVTS22 with that of Tn3 NM resolvase and Tn3 NM ZFR on their own substrates (nicked). The uncut reaction products shown in Figure 4.23 were nicked using the enzyme Nt.BstNBI. This is a lower percentage gel (0.7%) and it was run at 20V for 18 hours. Results are as described in Figure 4.23. NM resolvase and Tn3 NMZ yield significantly more resolution products than TALER6. Although they all show substrate depletion to the same extent, the accumulation of single-site cleavage products is characteristic of TALER6 activity. These reactions were carried out under standard recombination reaction (Section 2.16).

4.2.4 Binding Properties of TALER variants

The binding properties of TALERs were characterized using electrophoretic mobility bandshift assays as described in Section 2.15. These assays were carried out in TB buffer using 5% polyacrylamide gels supplemented with or without 10% glycerol. Glycerol is added to the gel mixture before polymerization to stabilize protein-DNA complexes (Sidorova *et al.*, 2010). A 78-bp Cy5-labelled dsDNA (Cy5-78) carrying a centrally placed 22-bp spaced T-site was generated by annealing a Cy5-labelled oligonucleotide to its complementary unlabelled bottom strand (Fig. 4.25).

The assignment of TALER complexes is provisional and is based on comparison with the behaviour of corresponding resolvases and ZFRs. TALER6 forms what appears to be synaptic complexes on Cy5-78 (Fig. 4.25). Increasing the protein concentrations leads from principally monomeric complexes at 200 nM to a mixture of synaptic complex and dimeric complex at 1600 nM. Increasing the protein concentration to 3200 nM leads to an aggregate formed at the well that does not migrate into the gel. The synaptic complex seems to dissociate into dimeric complexes during the run, indicating the instability of the complexes in the gel.

The binding properties of other catalytic domain variants tested in Section 4.2.2 were also analysed on Cy5-78 (5% PAGE + 10% glycerol) (Fig. 4.26). The bandshift patterns of TALER6 was as previously described. TALER6-WTR forms stable synaptic complexes under these conditions. This contrasts with previous reports on the inability of wild-type resolvase to form stable synaptic complexes with Tn3 *res* site I on bandshift assays. This TALER6-WTR synaptic complex with the DNA was also found to be stable in a 5% PAGE run in the absence of 10% glycerol (data not shown). TALER6-SY and TALER6-M however do not form synaptic complexes here but had bands that correspond to the dimeric complex (Fig. 4.26). Progression from monomeric complex as concentration of protein increased seemed slowest from TALER6 as the other catalytic domain variants had higher order complexes formed from 400 nM while all TALER6 monomeric complexes only

converted at 1600 nM. The differing abilities of these TALER proteins to synapse also indicates that complex formation is dependent on both the DBD and the catalytic domain. An attempt to detect different-sized complexes was made by using unlabelled dsDNA of 157 bp (UL-157) at equal ratio with Cy5-78, however, it seems that the size of the protein-DNA complexes attenuates any minor difference in DNA size (data not shown).

Analysing the effect of N-terminal truncations on TALER binding showed that the $\Delta 221$ TALER variant showed almost no observable complexes on the gel across the concentration ranges considered (Fig. 4.27). $\Delta 153$ and $\Delta 148$ had a similar binding pattern although $\Delta 148$ seemed to have a slightly higher binding affinity than $\Delta 153$. With $\Delta 119$, only monomeric complexes were observed at 200 and 400 nM and some synaptic complexes were observed at 800 nM. The differences in TALER binding ability based on the N-terminal truncation suggest that the varying recombination efficiency of these variants is dependent on the ability of the protein to bind as opposed to the placement of the catalytic domain relative to the crossover site. These results also provide an assurance that TALER binding activity observed here is based on complex formation with the DNA and not just multimerization due to disulphide bridge formation.

CY5-78 sequence

●GCTAGCAGTCAGATGTGGAACGGAAGAGCTATAAATTTATAATATTTGCTTCTTCCGTTTCCACATCTGAGCTCCCGG
 CGATCGTCAGTCTACACCTTTGCCTTCTCGATATTAATAATATTATAAACGAAGAAGGCAAAGGTGTAGACTCGAGGGCC

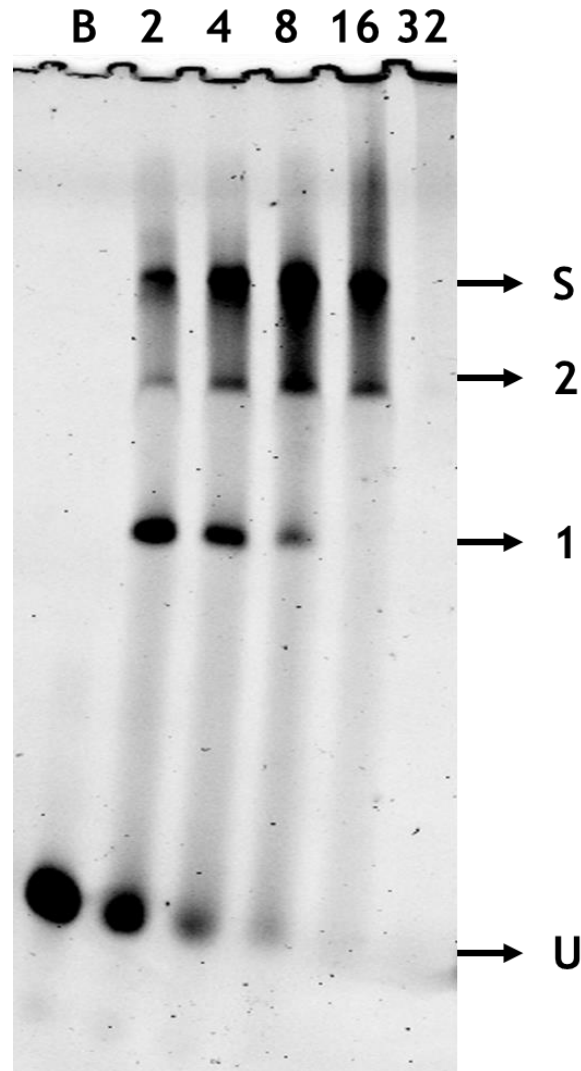


Figure 4.25: *In vitro* binding bandshift assay of TALER6 on Cy5-labelled 78-bp dsDNA with a T-site 22 architecture. The sequence of the 5' Cy5-labelled 78bp dsDNA (Cy5-78) is shown at the top with the location of the Cy5 dye indicated with a red circle. On a 5% polyacrylamide gel (with 10% glycerol) with TB buffer, TALER6 shows the formation of complexes like the monomeric, dimeric and synaptic complexes detected in previously described bandshift assays of activated mutants of Tn3 resolvase (Olorunniji *et al.*, 2008). Band assignment is based on this and the abbreviations are as follows: “S” (synaptic complex), “2” (dimeric complex substrate), “1” (monomer complex), and “U” (unbound labelled DNA). The DNA concentration was kept constant at 50 nM (Section 2.15) and protein concentrations (*100 nm) are indicated at the top of the gel. Lane annotated B (blank) contains no protein. Complex formation increased with increasing protein concentration.

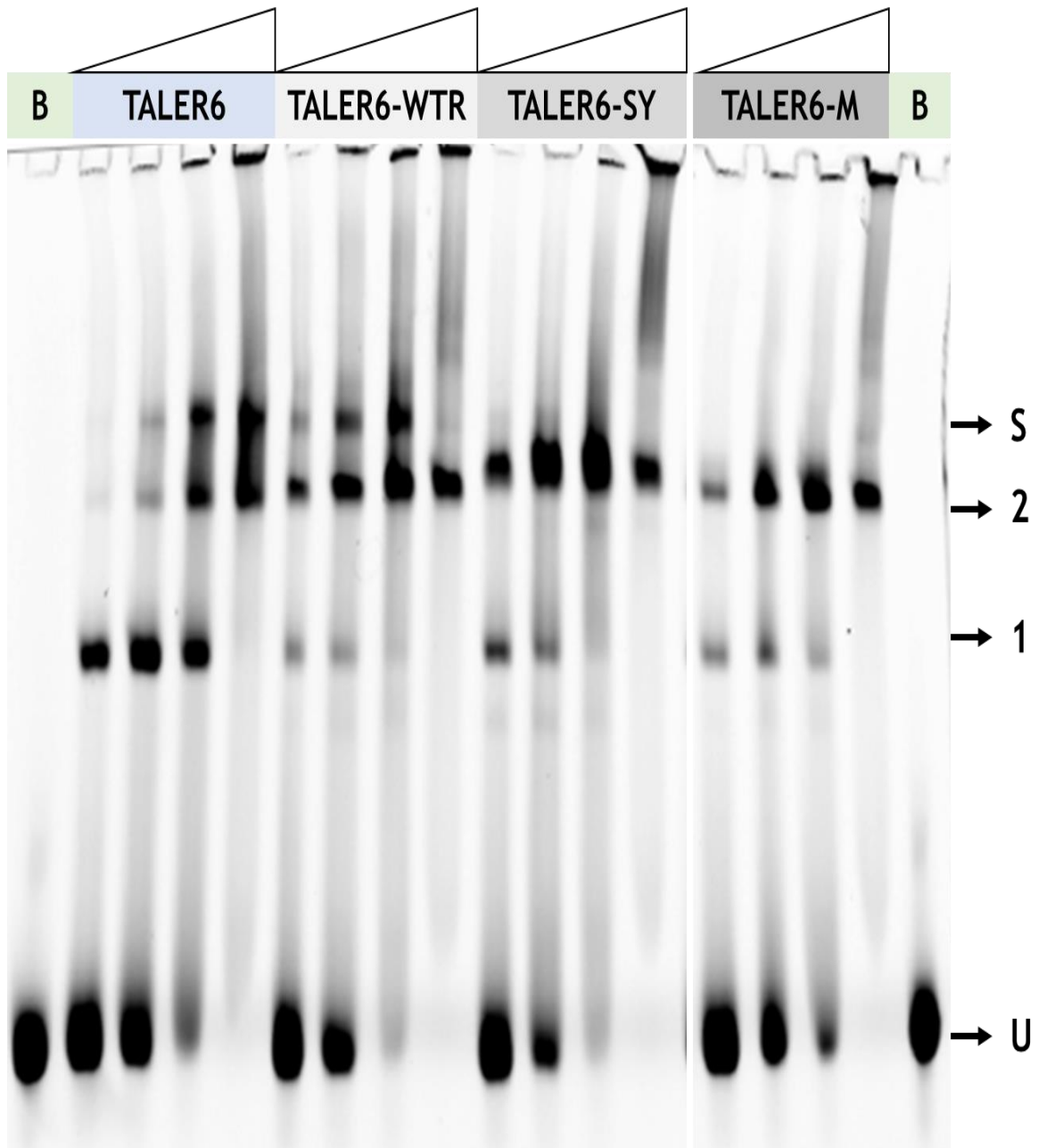


Figure 4.26: *In vitro* binding bandshift assay of catalytic domain variants of TALER6 on Cy5-labelled 78-bp dsDNA with a T-site 22 architecture. On a 5% polyacrylamide gel (with 10% glycerol) synaptic complexes are observed with TALER6 as before. TALER6-WTR, TALER6-SY and TALER6-M seem to form more complexes than TALER6 at lower protein concentration. However, TALER6-SY and TALER6-M do not show yield stable synaptic complex bands.

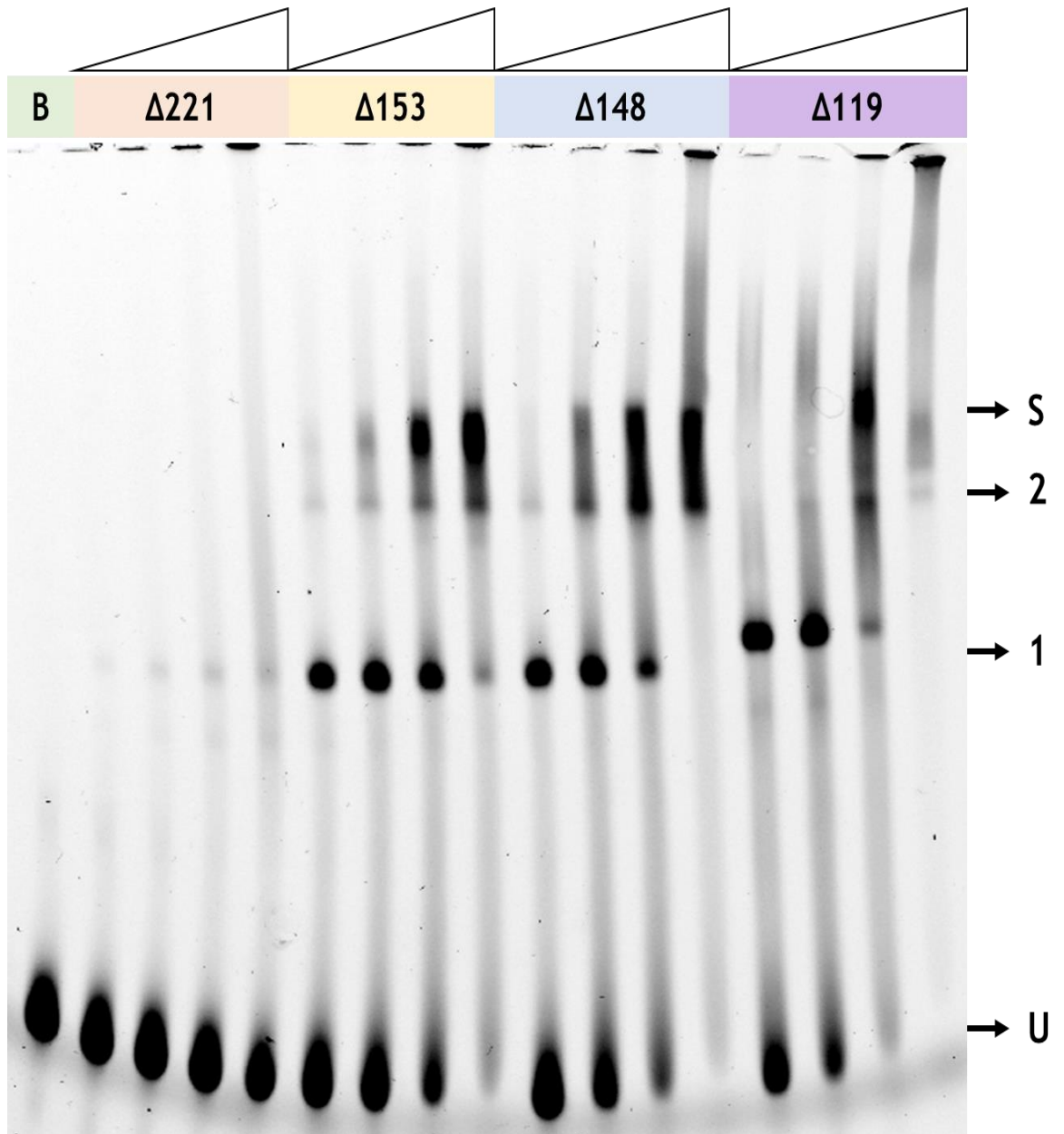


Figure 4.27: *In vitro* binding bandshift assay of N-terminal truncation TALER variants on Cy5-labelled 78-bp dsDNA with a T-site 22 architecture. On a 5% polyacrylamide gel (with 10% glycerol) synaptic complexes are observed with $\Delta 153$, $\Delta 148$, and $\Delta 119$ TALER variants. The $\Delta 148$ TALER variant used here is TALER6. Shorter truncation variants seem to have a higher affinity for Cy5-78. Very minimal complexes are observed with $\Delta 221$ TALER.

4.3 Discussion

4.3.1 *In vivo* activity of TALERs

While *E. coli* strains have been used widely for the cloning, expression and purification of TALE and TALE-derived proteins, it has been difficult to find published work where the *in vivo* activity of TALE proteins was assayed in *E. coli*. Niu *et al.*, 2015 and Kusano *et al.*, 2016 report on the use of TALENs in *Bacillus nematocida* B16 and *E. coli* respectively. They both used an assay that detected luciferase activity based on single-strand annealing repair of TALEN-mediated cleavage sites. However, the TALEN activities reported in both cases were quite low. Although Niu *et al.*, 2015 predicted that this could be due to lower transformation efficiencies, an alternative hypothesis could be introduced.

Bacterial genomes are exposed to TALE proteins and since the mechanism of TALE DNA-binding and target site recognition involves facilitated diffusion across the DNA, the presence of intracellular TALE proteins could disrupt cellular activities by blocking transcriptional start sites and essential regulatory elements. A consequence of this could be the trapping of TALE and TALE-based proteins in inclusion bodies to protect the genome, reducing protein availability, in this case for recombination catalysis. The insolubility of TALER proteins was consistently observed in this work during the purification of multiple TALER variants, giving some credence to this hypothesis. Since lower copy number expression vectors are used for MacConkey agar recombination assays, intracellular TALER availability might be lower than envisaged. This could also explain why a few colonies appear white in the MacConkey assay; cellular vulnerability or tolerance might be at play in these few cases. Although, the presence of long inverted repeats on the substrate plasmids might lead to homologous recombination and this might be a better explanation for the white colonies.

The implications of these findings for the work of Mercer *et al.*, 2012 could also be significant as the TALER design, target site definition and truncation studies they reported were largely done in *E. coli*.

4.3.2 Structural definition of TALER architecture

In agreement with previously published work on defining TALEN scaffolds, longer truncations of the N-terminus of the TALE DBD impeded DNA binding in the TALER constructs (Cuculis *et al.*, 2015; Gao *et al.*, 2012). This is in slight contrast to reports by Holt (2014), where it seemed that TALER constructs with fewer truncations than in the $\Delta 148$ TALER (such as $\Delta 119$ TALER) did not form the predicted synaptic complex. The $\Delta 119$ TALER, which is the longest variant (having the shortest truncation) analysed here, shows stronger DNA-binding properties than TALER6 ($\Delta 149$ TALER) and what appears to be a synaptic complex in the bandshift assay.

It is clearer now that the N-terminal TALE truncation point is highly important for TALER design and *in vitro* recombination. Of all the architectures observed, the $\Delta 148$ truncation still seems to be the best. Gao *et al.*, (2012) defined the crystal structure of a TALE variant (dTAL2) that shows that the region right after this (from residue 162 to 288) contained four cryptic repeats that had significant DNA-binding properties and that each cryptic repeat had similar structures to the repeats of the CRD (Section 1.3.1.2) (Fig. 4.28). These cryptic repeats are thought to be the initiators of TALE binding as they can bind to DNA irrespective of the CRD while the CRD cannot independently bind to DNA without this N-terminal region (Cuculis *et al.*, 2015; Gao *et al.*, 2012).

A putative DNA-binding residue at position 156 (R156) has also been implicated in direct DNA interaction (Gao *et al.*, 2015). The proximity of the $\Delta 153$ truncation to this residue might explain the reduction of recombination activity observed with the $\Delta 153$ TALER variant. The binding affinity of this TALER variant is slightly reduced compared to that of the $\Delta 148$ variant (TALER6) and might point to the destabilization of an interaction at R156. According to Szurek *et al.*, (2002) N-terminal truncations up to 153-aa abolish the secretion and translocation of TALE proteins via the Type III protein secretion system. This and the consistent use of the $\Delta 153$ scaffold in TALEN genome editing suggest that the residues after position 153 might be necessary for TALE protein binding activity.

It is possible that the additional residues (PAAQV) in the $\Delta 148$ TALER variant serve as a suitable structural linker that functionally separates TALE binding activity from the attached resolvase. This could explain why the length of the designed flexible linkers between the two domains did not yield significantly differential activities. The tested rigid linker might have reduced recombination activity due to increased stiffness between the domains, displaying the catalytic domain in a way that reduces its interaction with the crossover site.

4.3.3 T-site architecture

In contrast to the findings of Mercer *et al.* (2012), longer central DNA spacers between the two TALE DBD target sites seemed to be unsuitable for TALER activity. The longer TALER variant here, $\Delta 119$ which is similar to their $\Delta 120$ truncation variant did not show significant activity on T-sites with 26-bp and 28-bp spacers. This incongruity in the results could be a function of the different catalytic domain used in their work (Gin invertase). It could be also that their reported activity is a relic of the *E. coli in vivo* recombination platform used. The optimal spacer length identified here is consistent with the report by Jiullerat *et al.* (2014) that when the Fok1 nuclease domain is fused to the N-terminal end of the TALE DBD (the architecture used for TALER design), the active spacer length for cleavage activity falls within a narrow margin of 22-27 bp. Future work with binding and synapsis assays using labelled dsDNA of varying spacer lengths might help to elucidate the stage(s) of recombination at which spacer length discrimination comes into play.

T-sites with the 22-bp spacer length seemed to be the most preferred target although T-site 24 comes in close. This is similar to reports by Akopian *et al.* (2003) and Prorocic *et al.* (2011) in defining optimal ZFR target sites. The cut-off point of the resolvase catalytic domain might have a slight implication for TALER activity. TALER6X (cut-off point R144) shows a minor preference for T-site 24 while the longer TALER6 (cut-off point R148) prefers the shorter T-site 22. This could be a pointer to the importance of DNA turning and resolvase catalytic domain plasticity in the design for further optimization of the TALER architecture. Protein-DNA

modelling might provide more information about this and the analysis of T-site spacers and TALERs with incremental removal of single nucleotides/residues respectively might provide an experimental insight into this interplay.

4.3.4 TALER catalytic domain as main driver of recombination

The analysis of catalytic domain variants of Tn3 resolvase showed that the most hyperactive mutant tested, NM resolvase, yielded the most significant recombination activity on the target site. This is similar to the results generated with *res* site I and Z-sites by Prorocic *et al.*, 2011 and Olorunniji *et al.*, 2008. Since the hyperactive resolvase mutants were selected in the context of the wild-type resolvase architecture (Olorunniji *et al.*, 2008; Burke *et al.*, 2004) (Section 1.8), it is possible that these activating mutations are still not optimal for TALER activity and that mutations need to be identified within the TALER architecture that stabilize the new conformations of TALER complexes for recombination.

4.3.5 TALER activity and recombination product distribution

Attempts to alter recombination activities of TALER by changing reaction conditions did not prove fruitful. The current challenge with the most active TALER variant is the almost complete loss of topological selectivity. This allows for simultaneous catalysis of excision, integration and inversion. The propensity of TALERs towards intermolecular recombination also leads to an accumulation of linear cleaved products. Since supercoiling seems to drive the recombination reaction forward, linear products tend to become dead-end recombination products. While the analysis of TALER activity on single-site substrates simplifies the possibilities of recombination and shows clear integration, the focus of this work is excision which is dependent on the presence of the target site in direct repeat on the same molecule.

At the current stage, Tn3 NM TALERs do not carry out excision with the same fidelity as Tn3 NM resolvase or Tn3 NM ZFR. Many speculations about the reason for this have already been discussed. In addition, the binding and wrapping of the

TALE domain to a total of 34 bp flanking the central catalytic domain target site might lead to changes in the shape and structure of the DNA. This might prevent or reduce resolvase activity or the stability of the phosphoserine bond after cleavage, leading to the dissociation of TALER from the cleaved DNA.

One strategy that has not been considered in this work is the reduction of the number of TALE repeats within the TALE CRD used in the targeting the TALER. Boch *et al.*, 2009 reported that 10.5 or more TALE repeats provided robust gene activation. A shorter TALE variant that targets a T-site similar in length to the standard Z-site architecture, with 9-bp flanking the central 22-bp on either side, might actually yield considerable results. Less torsion might be applied on the DNA and/or the resolvase catalytic domain.

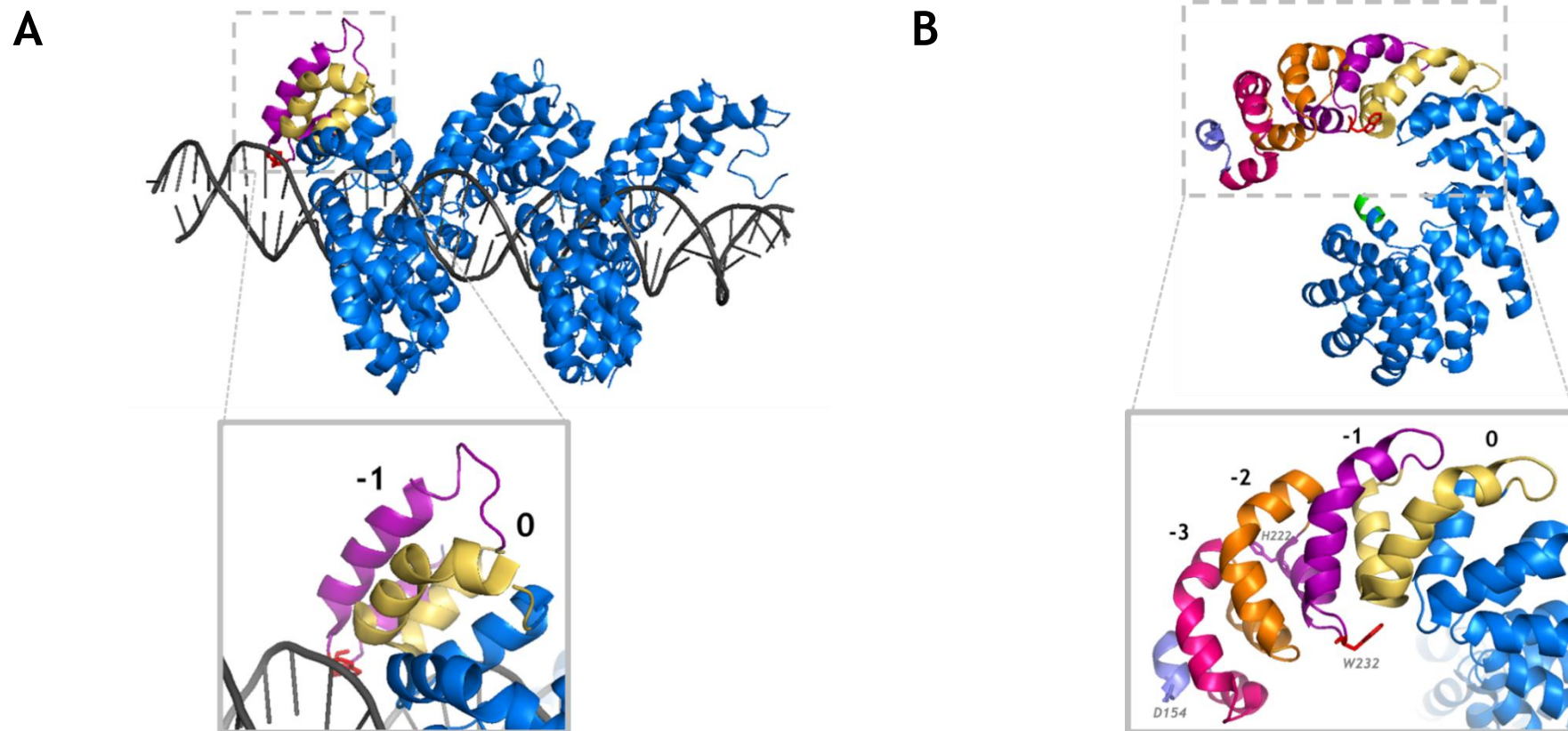


Figure 4.28: Crystal structures of PthXo1 and dTALE2 showing cryptic repeats in TALE NTR. A. 3UGM (Mak *et al.*, 2012) B. 4HPZ (Gao *et al.*, 2012). The N-terminal region of TALE proteins features 4 cryptic repeats that interact with DNA and are predicted to serve as anchor points for DNA-binding. The repeats are numbered -3, -2, -1 and 0. The residue W232 implicated in direct interaction with the 5' thymine at position 0 is shown as a red stick. D154 and H222, the starting TALE residues in the $\Delta 153$ and $\Delta 221$ TALER variants used in this work are also shown as sticks to indicate their position on 4HPZ.

Chapter Five: Construction and Characterization of TATA-TALER

5.1 Introduction

In Chapter 3, the design of a catalytic domain that targets the central 16 bp of the HIV CR_TATA_target sequence was reported. The mutations present in this Tn3 NM resolvase-based catalytic domain are Q13R, I77L, V107F, R120L, E132A, and I138V (Section 3.3.8). The selected mutant ZFR (ZR045) yielded specific and stable expected recombination products on the target mock-HIV substrate plasmid, pJU003 (HIV-ZLR). Significant *in vivo* recombination activity of TALERs has not been observed in the DS941 *E. coli* strain used in this work (Section 4.2.1). This led to the design and characterization of the current Tn3 Resolvase TALER architecture using *in vitro* techniques (Section 4.2). The selected TALER monomer architecture has a TALE DBD with 148 amino acids removed from its N-terminal region, a central repeat domain that targets an 18-bp sequence, and 63 amino acids retained in its C-terminal region ($\Delta 148/+63$) (Section 4.3.2). This TALE DBD is fused to the C-terminal end of the Tn3 NM resolvase-based catalytic domain (cut off at residue 148) using a 6-aa flexible linker (GSGGSG) followed by a SpeI restriction site (encoding TS). Similar to the optimal ZFR target site (Z-site) reported by Prorocic *et al.* (2011) and Akopian *et al.* (2003), the optimum spacer length of the target Tn3 resolvase based-TALER target site (T-site) identified was 22 bp. These observations laid the foundation of the design of the full HIV-targeting TALERs reported in this Chapter.

5.2 Results

5.2.1 *In vitro* analysis of the activity of the selected mutant catalytic domain

To observe the *in vitro* activity of ZR045, the coding sequence of the catalytic domain of ZR045 was cloned from the *in vivo* low-level expression plasmid pJUM045 into a high-level overexpression plasmid backbone to generate pJUM245 as described in Section 2.10.1. High copy number *in vitro* recombination substrate plasmids carrying two copies of the HIV-ZLL, HIV-ZRR and HIV-ZLR Z-sites flanking a kanamycin gene were also designed and were called pJU201, pJU202 and pJU203 respectively (Section 2.6). ZR045 was purified using nickel-affinity chromatography (Section 2.12) and its activity was analysed on these plasmids (Fig.

5.1). Inversion, resolution and cleavage products were observed with pJU202 and pJU203, and only cleavage products were observed with pJU201. It is important to note that the amount of resolution products observed here was minimal.

It was necessary to analyse the activity of the catalytic domain from ZR045 within the context of the TALER architecture before designing the full HIV-TALER proteins. To do this, the catalytic domain coding sequence from pJUM045 was swapped with that of Tn3 NM resolvase in pJUM501 using the NdeI/SpeI cloning strategy (Section 2.10.1). This created a reading frame coding for a TALER (TALER245) with the ZR045 catalytic domain with the Δ 148 architecture. A new substrate plasmid, pJU515 (HLR-T1297) containing 22-bp spaced T-sites with the central 22-bp sequence of the HIV CR_TATA_target site flanked by TALE1297 DBD target sequences was generated as well (Fig. 5.2). The activity of the purified TALER245 on pJU515 is similar to that observed with ZR045 on pJU203, with observable but low-abundance resolution products and considerable cleavage products (Fig. 5.3).

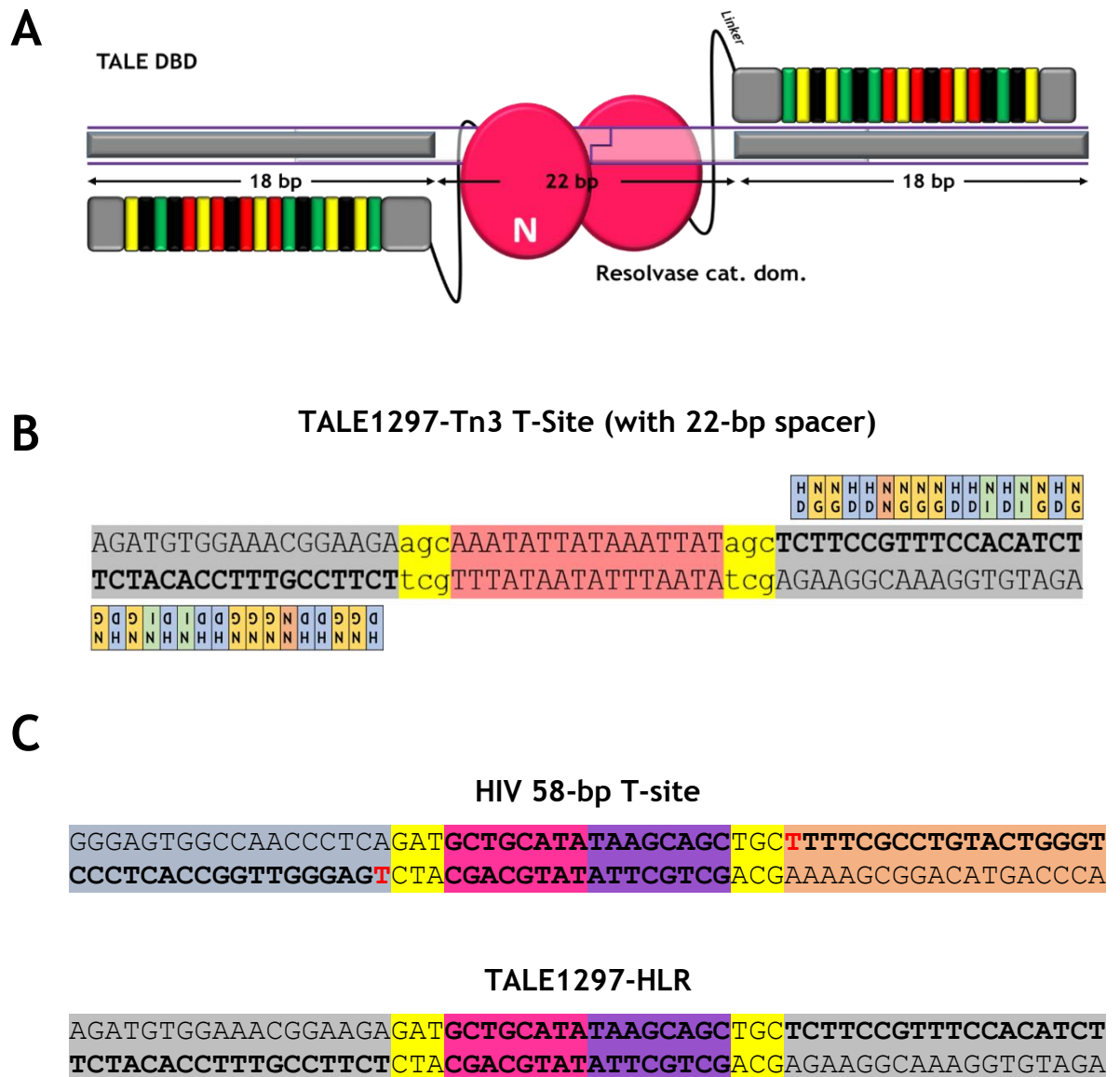


Figure 5.2: Design of new TALE1297/HIV T-site. A. Cartoons showing TALER monomers bound to a 22-bp spaced T-site architecture. B. The sequence of TALE1297-Tn3 T-site present in IVTS22 along with the TALE1297 RVDs is provided. The central 16-bp Tn3 *res* site is highlighted in red, the 3-bp spacer in yellow and the TALE1297 DBD target sequence in grey. C. The expected 58-bp T-site from the CR_TATA_target sequence shows the central 16 bp coloured according to Figure 3.4, with the left and right halves of the catalytic domain target site in pink and purple respectively. The binding sites for the left and right HIV TALE DBD are highlighted in blue-grey and orange. To generate the new TALE1297-HLR T-site, the target HIV TALE DBD sequences were swapped with TALE1297 DBD target sequence.

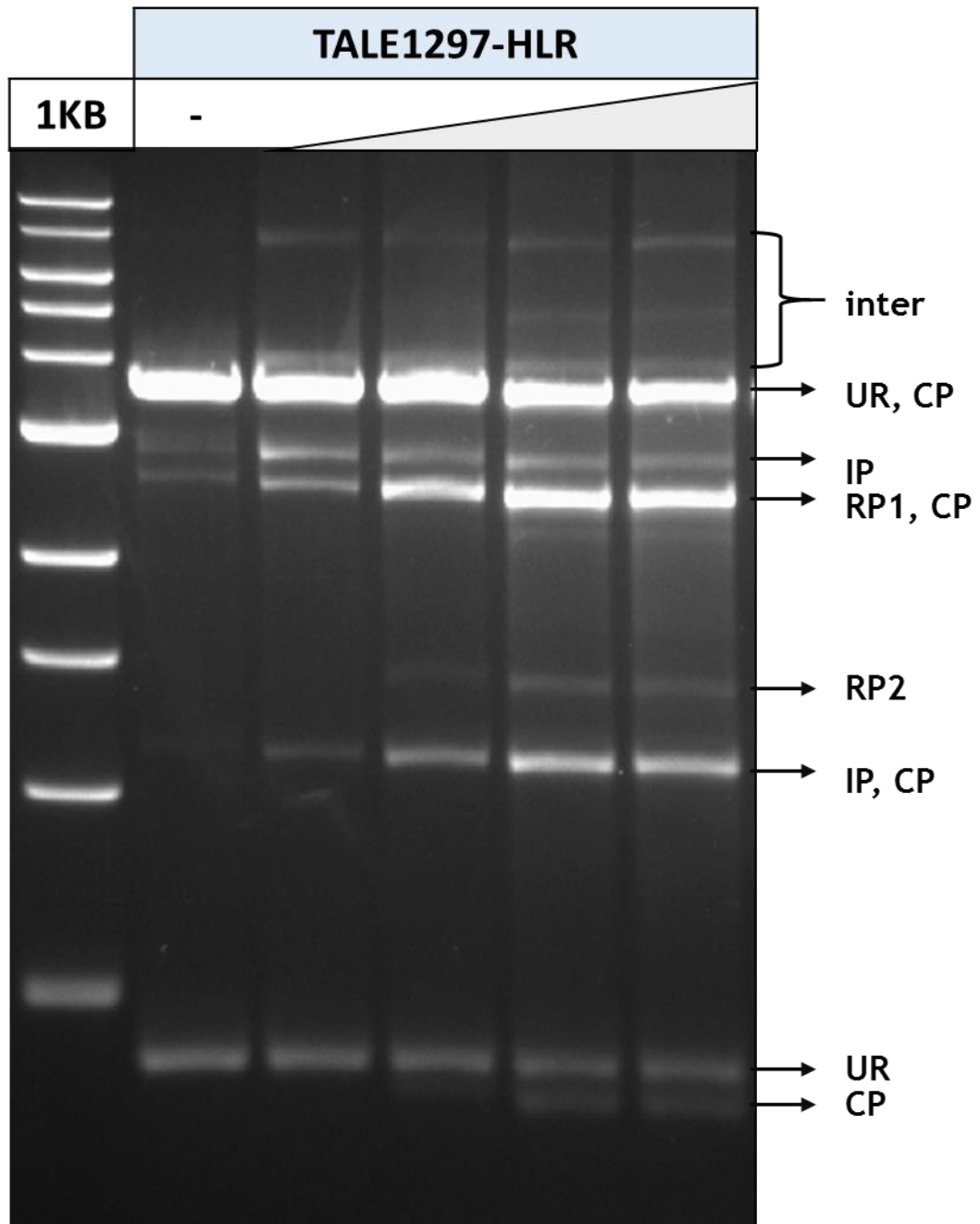


Figure 5.3: *In vitro* activities of TALER245. The activity of TALER245 on the TALE1297-HLR substrate (pJU515) with increasing protein concentration is shown. Product sizes are as indicated in Figure 2.10A. The smaller resolution product, RP2, is definitive of recombination activity. These reactions were carried out under standard recombination reaction conditions (Section 2.16).

5.2.2 Design of the HIV CR_TATA TALE DBDs

Up to this point in this work, T-sites have been designed to have a central spacer sequence (optimally 22 bp) flanked by the same TALE DBD target sequences on both sides. This is not a feature of the full HIV CR_TATA_Target. The sequences flanking the central 22 bp of the HIV CR_TATA_target site are very different. This meant that two different proteins had to be designed to direct the CR catalytic domain to the full site. The construction and design of the HIV TATA-TALE DBDs is illustrated in Figures 5.4, 5.5, 5.6 and 5.7. Two TALE-expression plasmid vectors (pHIV-TALEA and pHIV-TALEB) for targeting the left and right 18-bp HIV TALE binding sequences were ordered from Thermo Fisher Scientific (Section 2.5). Multiple cloning strategies were utilized to generate the $\Delta 148/+63$ TALER versions of these proteins with the Tn3 NM resolvase catalytic domain fused to the N-terminal end as described in Figure 5.6. The resulting overexpression plasmids, pJUM602 and pJUM612 were used to express and purify the new TALER proteins, TALER100 (left-HIV-T-targeting) and TALER101 (right-HIV-T-targeting) respectively. The sequence of TALER100 is provided in Figure 5.8. The RVDs of TALER100 and TALER101 along with their target binding sites are shown in Figure 5.9.

Three *in vitro* recombination TALER substrate plasmids were designed featuring T-sites with the standard Tn3 *res* site | 22-bp spacer sequence flanked on the left/right by TALE DBD target sequences of TALER100/TALER100 (HTA-Tn3-HTA), TALER100/TALER101 (HTA-Tn3-HTB) and TALER101/TALER101 (HTB-Tn3-HTB) (Figure 5.10). The plasmids were pJU504, pJU506 and pJU508 respectively (Section 2.6).

In vitro activity of TALER100 and TALER101 on these substrates demonstrated highly specific activity, with recombination products observed only when the appropriate protein targets its own substrate plasmid (Figure 5.11). With pJU506 (HTA-Tn3-HTB), a combination of TALER100 and TALER101 was required for recombination activity. The product distribution was similar to that observed with TALER6 on IVTS22 (Section 4.2.2).

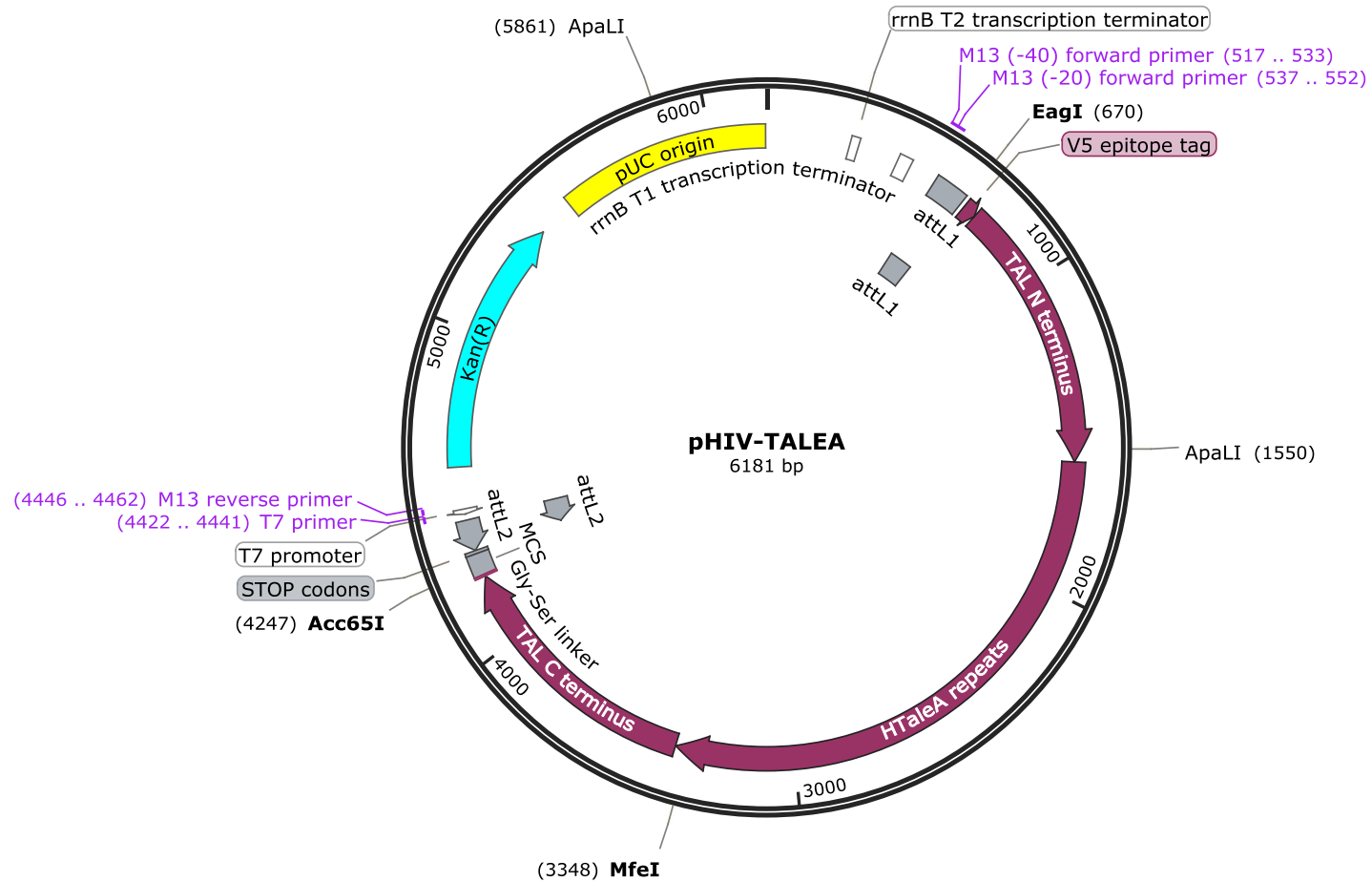


Figure 5.4: Plasmid map of GeneArt pPHIV-TALEA as supplied by Thermo Fisher Scientific. The TALE gene encoded contained the full N-terminus and C-terminus region. The restriction sites used in the generation of the TALER variant are indicated.

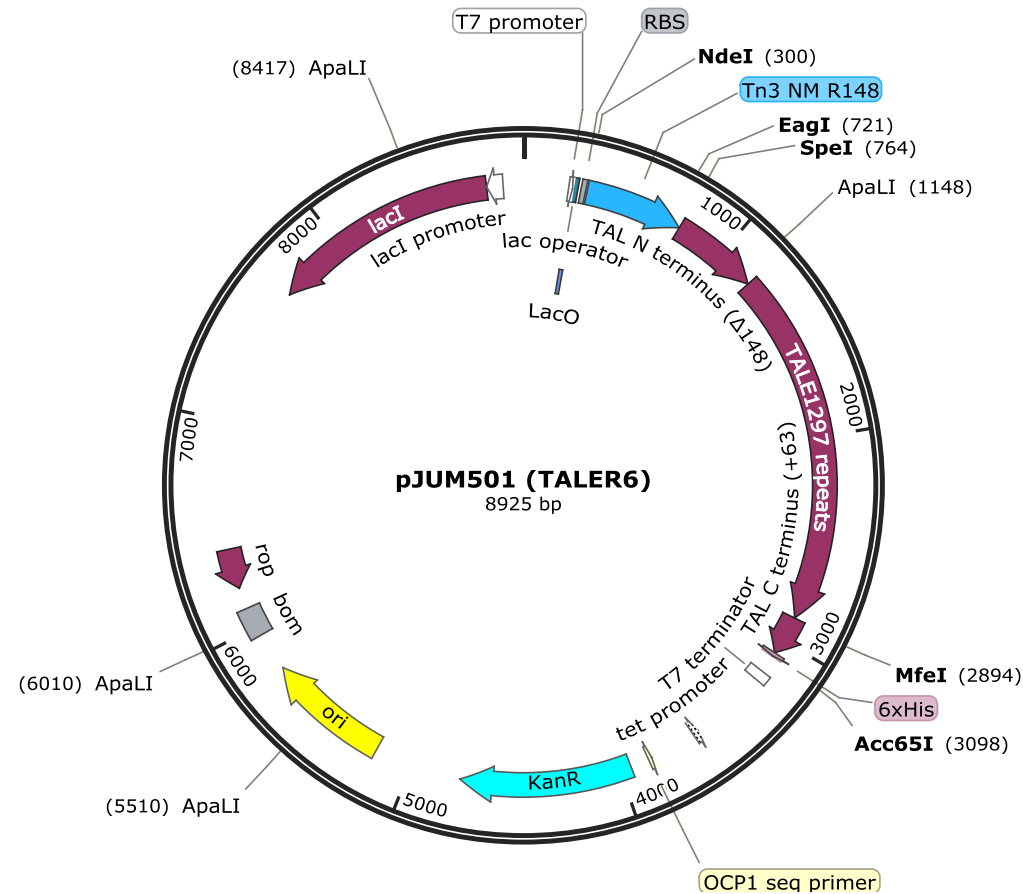
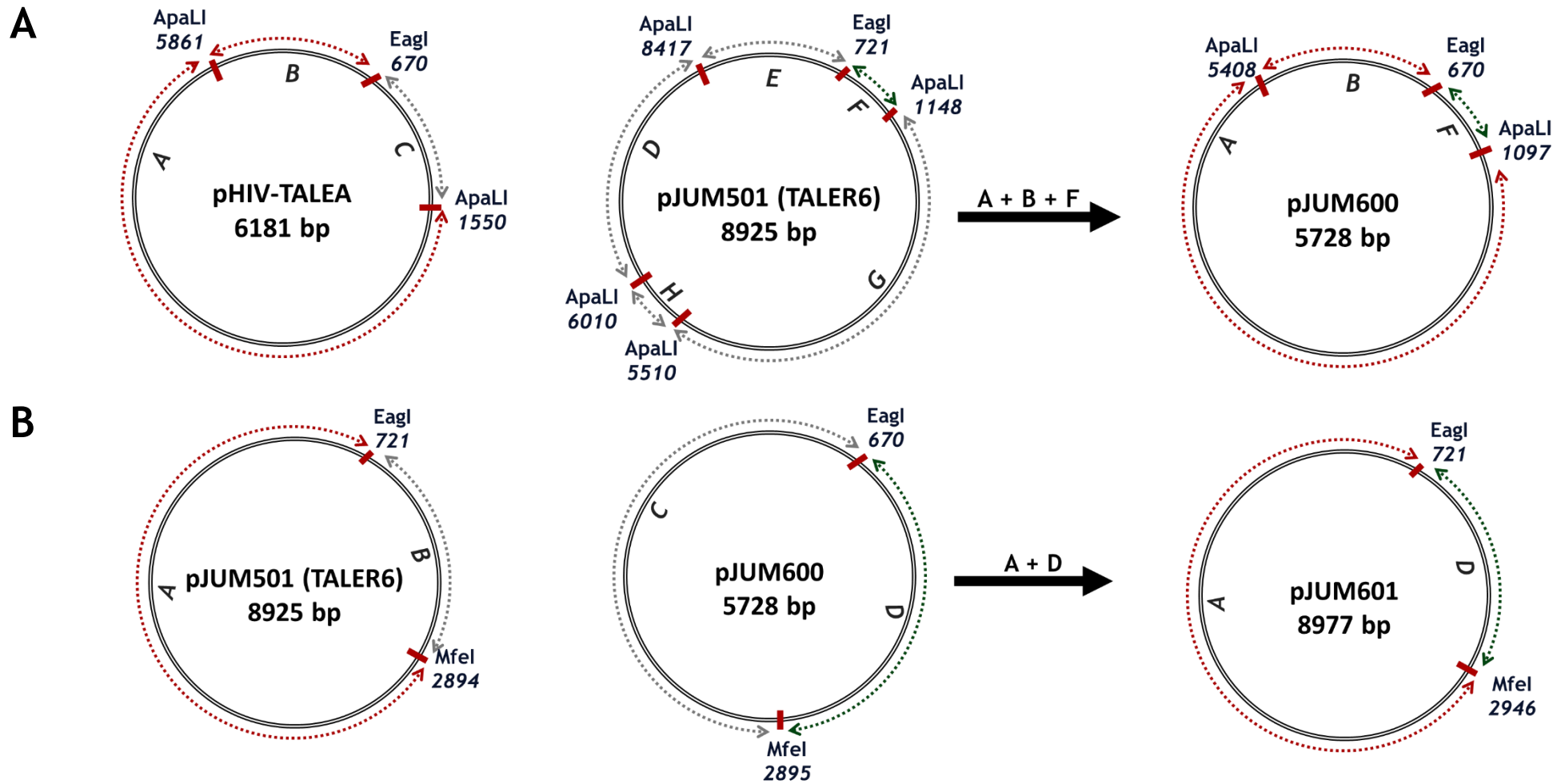


Figure 5.5: Plasmid map of pJUM501 (TALER6). The plasmid map of the overexpression plasmid of TALER6 is shown (Section 2.6). TALER6 has the $\Delta 148/+63$ TALER architecture (Table 4.6). The restriction sites used in the generation of the $\Delta 148/+63$ HIV-TALEA expression plasmid variant are indicated.



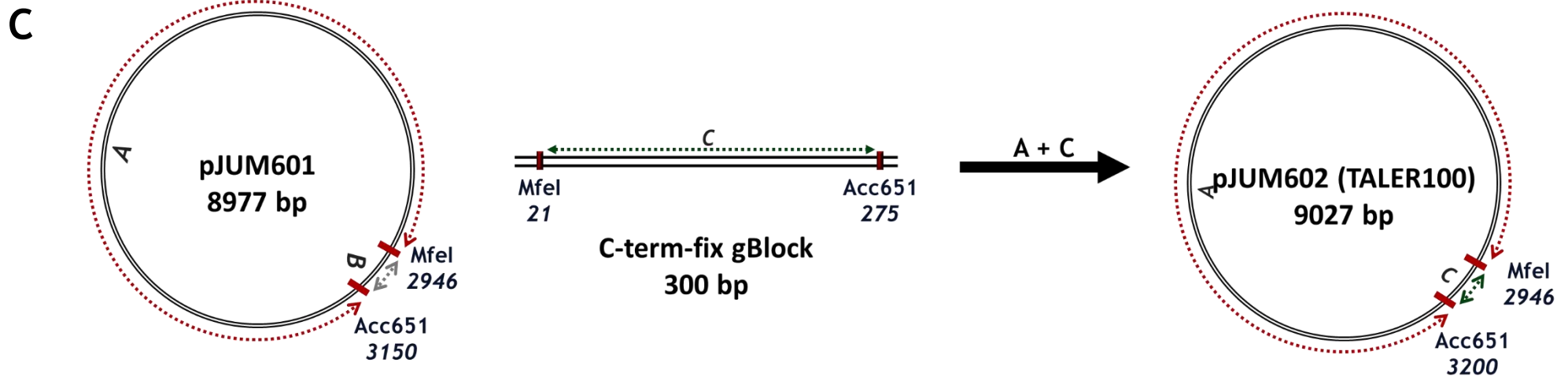


Figure 5.6: Schematic showing the cloning strategy for the construction of $\Delta 148/+63$ HIV-TALERA expression plasmid. A. Fragments from the ApaLI/EagI restriction digest of pHIV-TALEA and pJUM501 were ligated to generate pJUM600. This led to the introduction of the $\Delta 148$ N-terminal truncation into the HIV-TALEA coding region. B. The EagI/MfeI plasmid backbone of TALER6 (A) was ligated with the EagI/MfeI fragment (D) from pJUM600. This moved the carrying the TALE DBD ORF into the appropriate TALER overexpression plasmid backbone. C. The cloning in B led to the generation of a C-terminally truncated TALER variant with a frameshift mutation leading to a premature stop codon. To restore the C-terminal +63 architecture, a 300-bp double-stranded DNA fragment (CtermTalGcas: Section 2.5) was ordered from IDT was digested using MfeI and Acc651. The digested fragment was ligated into a MfeI/Acc651 digested pJUM601 plasmid backbone. This resulted in the generation of the $\Delta 148/+63$ HIV-TALERA expression plasmid, pJUM602. The encoded TALER, TALER100 targets the left TALE DBD target sequence of the full HIV T-site. The expression plasmid for TALER102 (pJUM612) was also generated from HIV-TALEB using the same cloning strategy from (as shown in A to C).

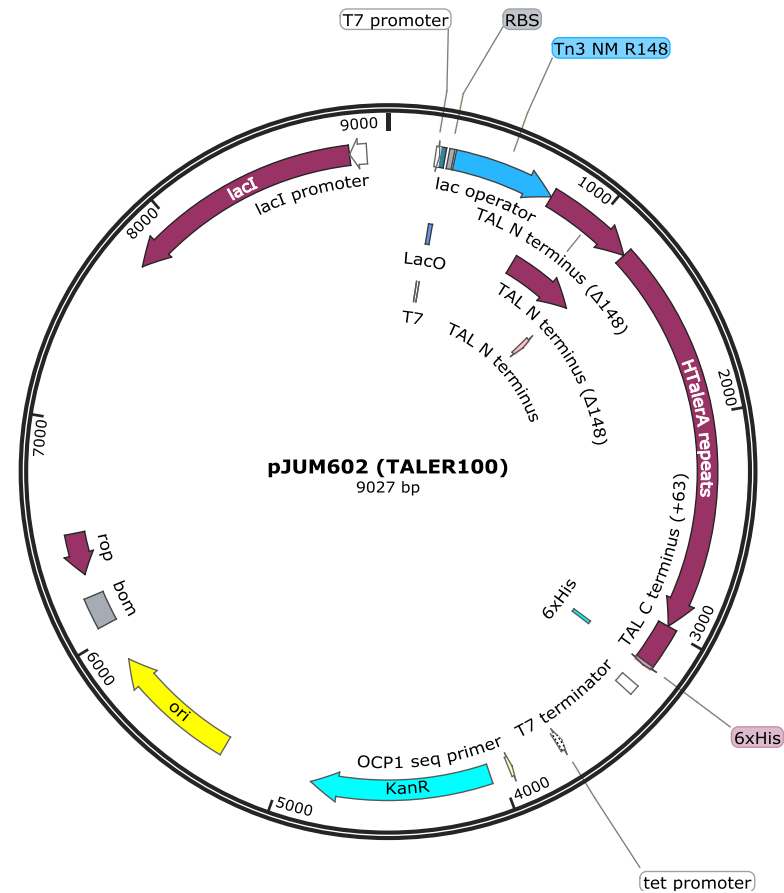


Figure 5.7: Plasmid map of pJUM602 (TALER100). The plasmid map of the overexpression plasmid of TALER100 is shown. TALER100 has the $\Delta 148/+63$ TALER architecture and targets the left TALE DBD target sequence of the full HIV T-site.

MAIFGYARVSTSQQSLDIQIRALKDAGVKANRIFTDKASGSSTDREGLDLLRMKVKEGDVILVKKLDRL
 GRDTADMIQLIKEFDAQGVAVRFIDDGI STDSYIGLMVVTILSAVAQAERRRILERTNEGRQEAKLKGI
 KFGRRRTVDRGSGGSGTSPAQAQVDLRTLGYSTQQQEKIKPKVVRSTVAQHHEALVGHGFTHAHIVALSQH
 PAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGG
 VTAVEAVHAWRNALTGAPLNLTPEQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGG
 KQALETVQRLLPVLCQAHGLTPEQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNNGGK
 QALETVQRLLPVLCQAHGLTPEQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNNGGKQ
 ALETVQRLLPVLCQAHGLTPEQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNNGGKQA
 LETVQRLLPVLCQAHGLTPEQVVAIASNNGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQAL
 ETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGKQALE
 TVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNNGGKQALET
 VQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETV
 QRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPQQVVAIASHDGGRPALESIV
 AQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAHHHHHH*

Figure 5.8: Sequence of TALER100. The colour scheme here is as defined in Figure 4.7. Tn3 NM resolvase (grey: cut off at residue 148) is linked to the N-terminal region of the TALE domain (truncated at residue 148) using a 6-aa GSGSG linker followed by a SpeI restriction site (TS). Residue 149 of the TALE DBD is underlined and in bold. The four cryptic repeats of the TALE NTR are also shown.

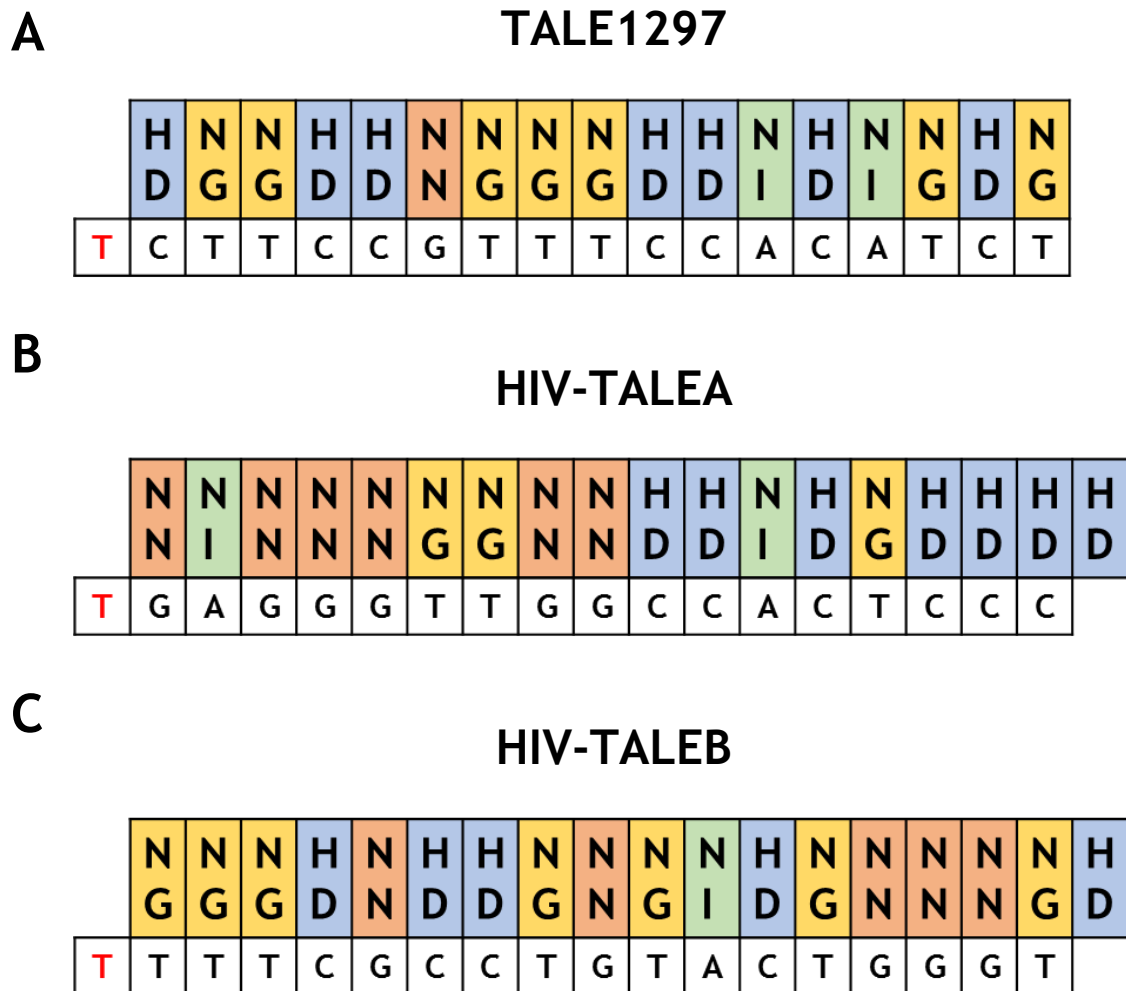


Figure 5.9: TALE RVDs and target sequences. The three TALE DBDs used in this work along with their 17-bp target sequences are shown. The 5' thymine is shown in red and the sequence reads from 5' to 3'. The RVDs are coloured as blue (HD), yellow (NG), orange (NN) and green (NI). In all three proteins, the RVD HD is associated with cytosine, NG with thymine, NN with guanine and NI with adenine. In the HIV-TALEs, an additional RVD, HD is present. This means the HIV T-site can be extended to 60 bp. The additional RVD is associated with the nucleotide present in the full CR_TATA_target sequence in the appropriate location.

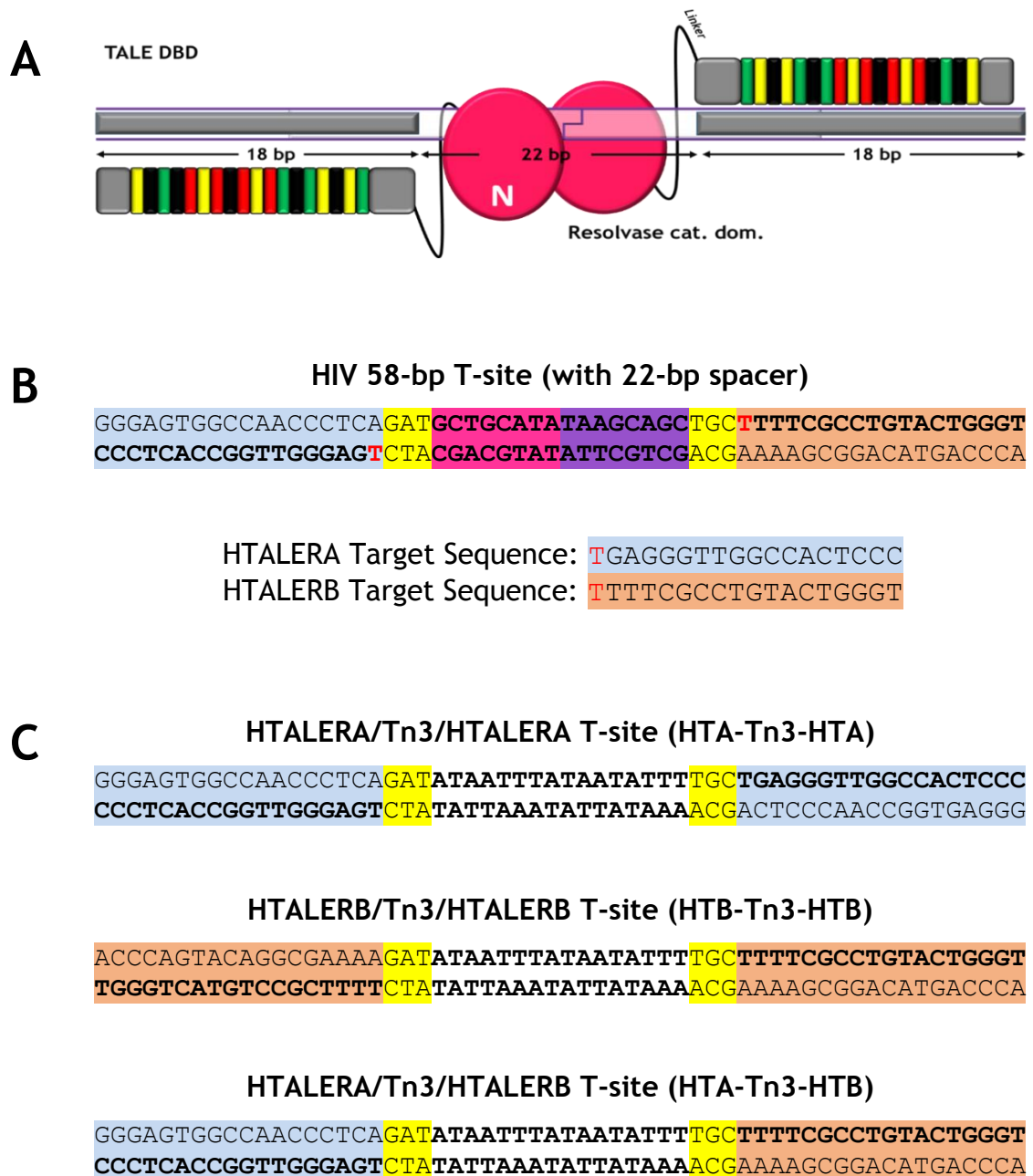


Figure 5.10: Design of new T-sites. A. Cartoon showing TALER monomers bound to a 22-bp spaced T-site architecture. B. The sequence of the 58-bp T-site from the CR_TATA_target sequence as coloured in Figure 5.2 is shown. The sequences of the left and right HIV TALE DBD are indicated and highlighted in blue-grey and orange. C. Three new T-sites with the central Tn3 16-bp *res* site I sequence flanked by Left/Left, Right/Right or Left/Right TALE DBD target sequences are shown.

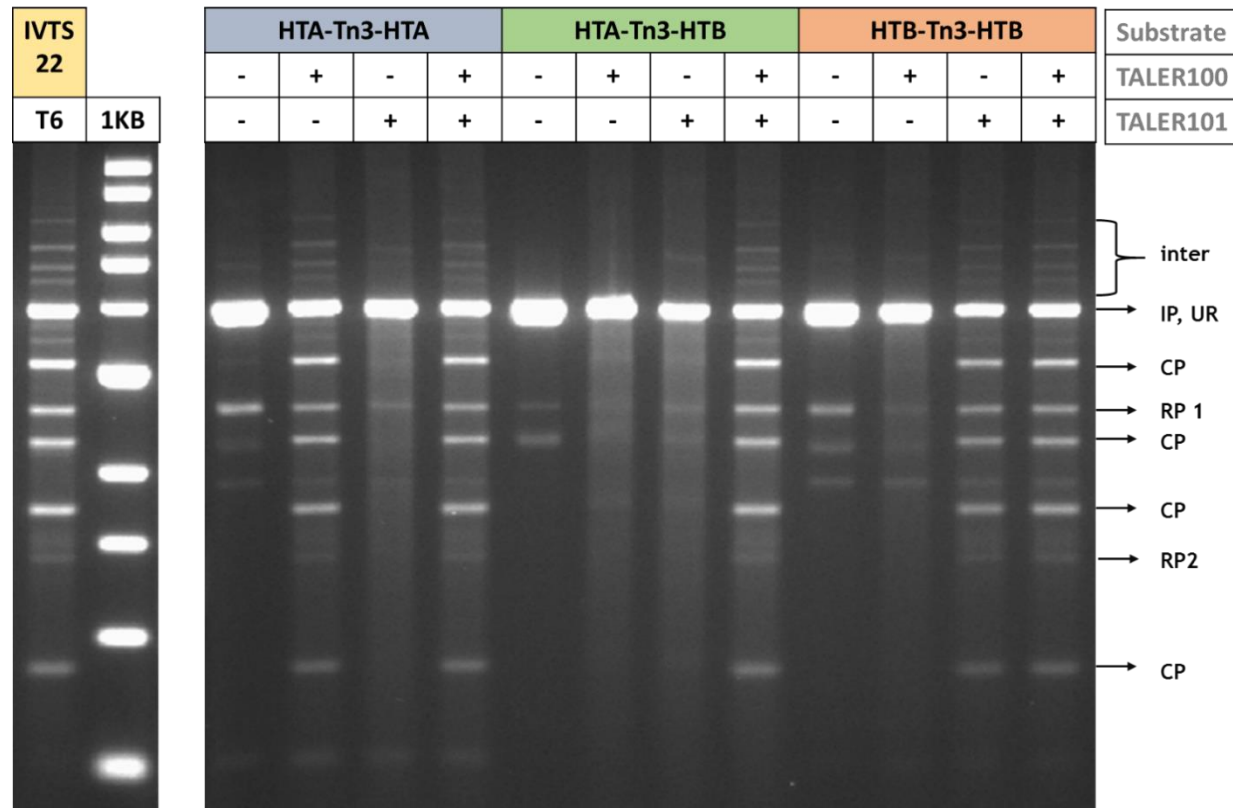


Figure 5.11: *In vitro* activities of TALER100 and TALER101 (AlwNI-digested). The reactions here contained either no protein, one of TALER100 or TALER101, or a mixture of the two proteins in 1:1 ratio. Final protein concentration in the reaction was kept constant at 400 nM and DNA concentration at 25 µg/ml. Product sizes are as indicated in Figure 2.10A. The smaller resolution product, RP2, is definitive of recombination activity. Extra bands in control reactions are unidentified. Strong orthogonality can be observed with each protein only active on its target substrate plasmid and significant recombination on HTA-Tn3-HTB only observable when both proteins are present.

5.2.3 Construction of the full HIV CR_TATA TALERS

As the catalytic domains and the DBDs targeting the HIV CR_TATA_target site had been analysed individually *in vitro*, the next step was to combine the components to generate TALERS that target the full sequence. Two different catalytic domains from ZFR mutants analysed in Chapter 3 were selected for this: the final HIV-targeting ZFR, ZR045 and an intermediary design mutant, ZR012. ZR012 (NM + K29E I77L V107F) has less mutations than ZR045 (NM + Q13R I77L V107F R120L E132A I138V) and it was predicted that targeting half of the full target site with a simpler mutant might tame the accumulation of cleavage products observed with the *in vitro* analysis of ZR045 and TALER245 (Section 5.2.1). The Tn3 NM resolvase catalytic domain sequence in pTALER100 and pTALER101 were swapped with NdeI/SpeI restriction fragments from pJUM012 and pJUM045 to generate pTALER102 (cat. dom. 012/TALE100), pTALER103 (cat. dom. 012/TALE101), pTALER104 (cat. dom. 045/TALE100) and pTALER105 (cat. dom. 045/TALE101).

An *in vivo* recombination substrate plasmid, pJU511 (HIV58T) with T-sites as the central 58-bp of the CR_TATA_target sites was also designed (Figure 5.10B).

5.2.4 Characterization of the activity of HIV CR_TATA TALERS

The four HIV-TALERS (TALER102, TALER103, TALER104 and TALER105) were purified and their *in vitro* recombination activities were analysed in combination with TALER100 and TALER101 on pJU504 (HTA-Tn3-HTA), pJU506 (HTA-Tn3-HTB), pJU508 (HTB-Tn3-HTB) and pJU511 (HIV58T) (Fig 5.12 and Fig. 5.13). The results from Figure 5.12 show an accumulation of cleavage products characteristic of TALER activity. With the single TALER analysis on HTA-Tn3-HTA and HTB-Tn3-HTB, TALER104 and TALER105 do not show significant resolution products on their respective substrates. However, it is clear that the TALE DBDs target the catalytic domains to their respective target sequences.

The analysis of undigested reaction products in Figure 5.14 showed that the activity of TALER102, TALER103, TALER104 and TALER105 yielded a significantly

higher amount of double-site cleavage products on HIV58T than on HTA-Tn3-HTB. It was surprising to observe that when TALER104 and TALER105, that have the most active catalytic domain, are complemented together, very minimal cleavage or resolution products is observed, yet topoisomerase activity is clearly demonstrated (Fig. 5.15). The TALERS with the Tn3 NM resolvase catalytic domain (TALER100 and TALER101), even when combined with TALER102, TALER103, TALER104 and TALER105, demonstrate no activity on HIV58T. Some resolution and inversion products can be detected with combinations of TALER102, TALER103, TALER104 and TALER105 on HIV58T.

As the levels resolution products were relatively low, the MacConkey agar-based assay (Section 2.11) was used to generate a colorimetric output of the *in vitro* recombination activity of the mutant TALERS on the HIV CR_TATA_target site. To do this, an *in vivo* TALER recombination substrate plasmid, pJU550 (HIV58T-L) carrying the 58-bp CR_TATA_target sequence as T-sites was designed. pJU550 DNA was purified in an amount sufficient for *in vitro* recombination analysis. Recombination reactions were carried out using HIV_TALERS (TALER102, TALER103, TALER104 and TALER105). The activity of purified Tn3 NM resolvase and TALER6 on their *in vivo* recombination substrate plasmids (pMP243 and TS22 respectively) served as controls. Figure 5.16 shows the results of the *in vitro* recombination assay. Again, very low levels of resolution products could be observed. Aliquots of the recombination reactions were ethanol-precipitated (Section 2.7.6). Approximately 200 ng of the recovered DNA was then transformed into *E. coli* DS941 electrocompetent cells. After expression, the transformed cells were plated on MacConkey agar plates (with 0.1% galactose and 50 µg /ml kanamycin) and incubated overnight. Overnight cultures were also set up with 1/100 dilutions of the expressed cells in L-broth (containing kanamycin) (Section 2.11).

Photographs of the plates after incubation reveal only a few transformants for the reactions with the HIV_TALERS, suggesting loss of transformable plasmids, perhaps due to substrate plasmid cleavage without ligation (Fig. 5.18). A few of the colonies were white. Interestingly this substrate depletion, to a lesser degree, is also evident with TALER6 on pTS22. However, in the surviving clones, cells

harbouring resolution products seem to present in a ratio of 1:1 with cells harbouring *galk*-containing plasmids. This 1:1 ratio is similar to the result with Tn3 resolvase on pMP243. The red colonies, which carry *galk*-containing plasmids probably contain inversion and intermolecular recombination products while the white colonies carry the resolution products where the *galk* gene has been excised. The analysis of the prepared DNA from the overnight liquid cultures show this as well (Fig. 5.17). Here, no plasmid DNA was recovered for TALER102/TALER103 activity on HIV58T-L, indicating almost complete substrate cleavage in the sample analysed.

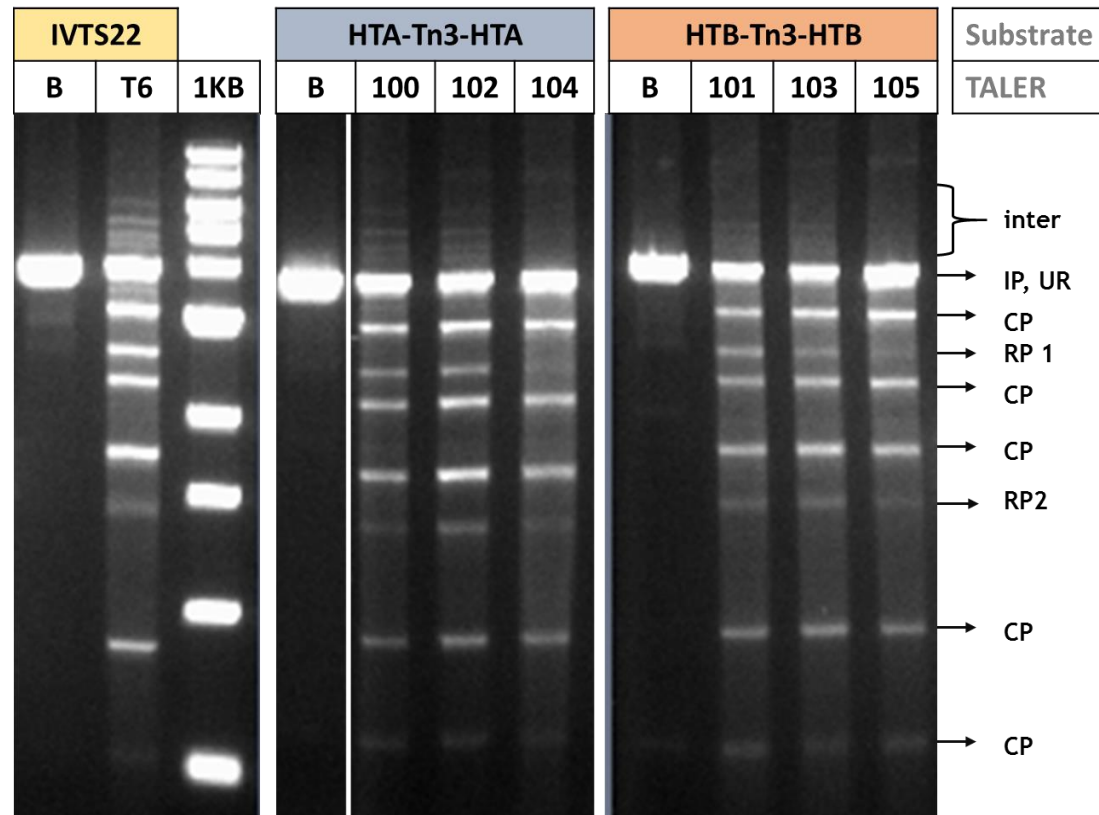


Figure 5.12: *In vitro* recombination activities of HIV TALERs on HTA-Tn3-HTA and HTB-Tn3-HTB substrate plasmids. The TALERs with the left HIV TALE DBD and Tn3 NM, ZR012 and ZR045 cat. doms. (TALER100, TALER102 and TALER104 respectively) were tested on HTA-Tn3-HTA while the TALERs with the right HIV TALE DBD and Tn3 NM, ZR012 and ZR045 cat. doms. (TALER101, TALER103, and TALER105 respectively) were tested on HTB-Tn3-HTB. The activity of TALER6 on IVTS22 serves as a control. AlwNI digest was used to show product distribution. Recombination product sizes are as indicated in Figure 2.10C. The bigger resolution product, RP1, is definitive of recombination activity. The least amount of resolution products is observed with TALER104 and TALER105.

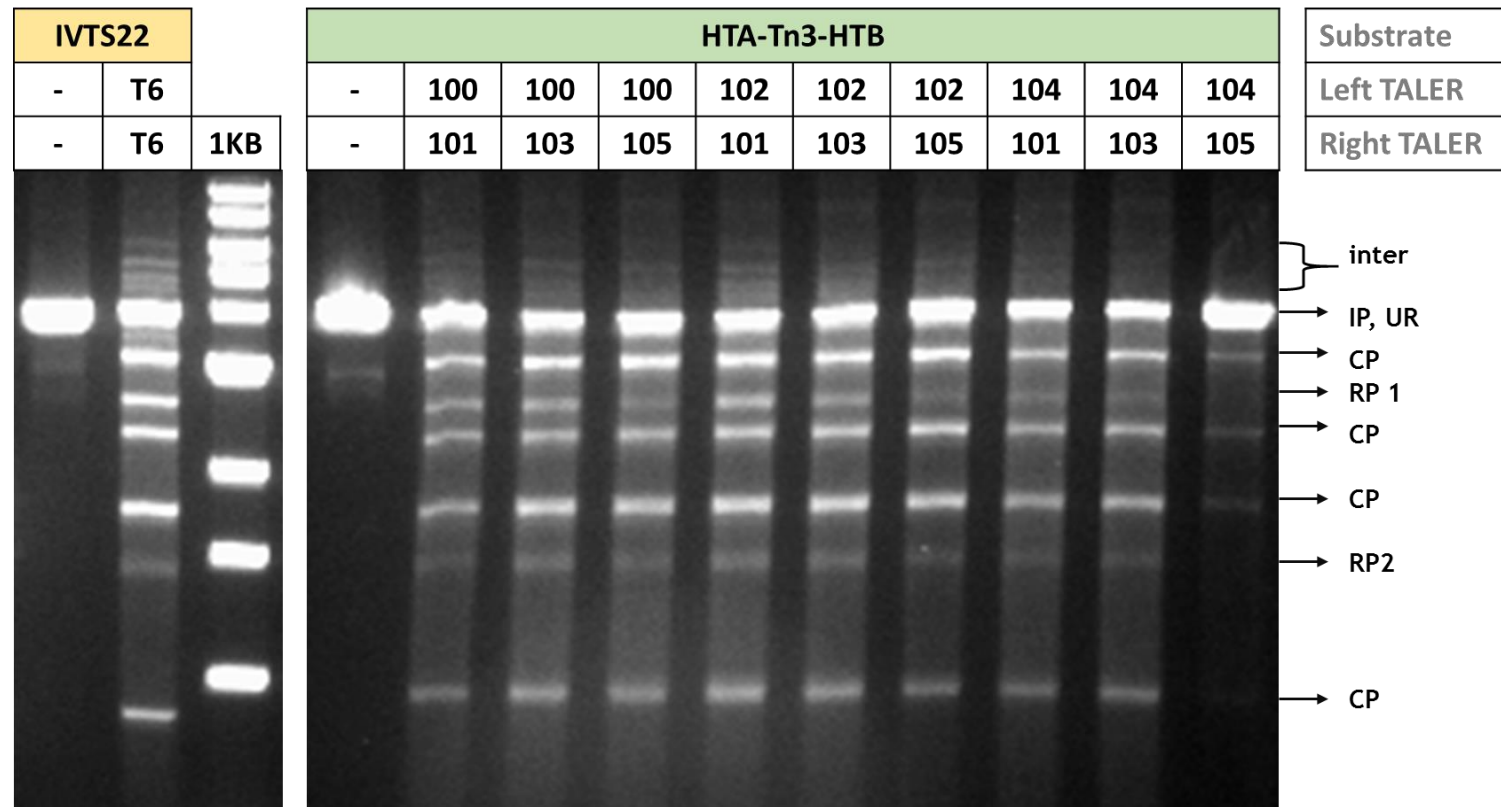


Figure 5.13: *In vitro* recombination activities of HIV TALERs on HTA-Tn3-HTB substrate plasmid. The reactions here contained a mixture of two proteins in 1:1 ratio. Final protein concentration in the reaction was kept constant at 400 nM and DNA concentration at 25 µg/ml. The activity of TALER6 on IVTS22 serves as a control. AlwNI digest was used to show product distribution and product sizes are as indicated in Figure 2.10C. The bigger resolution product, RP1, is definitive of recombination activity. The least amount of resolution products is observed with TALER104 and TALER105.

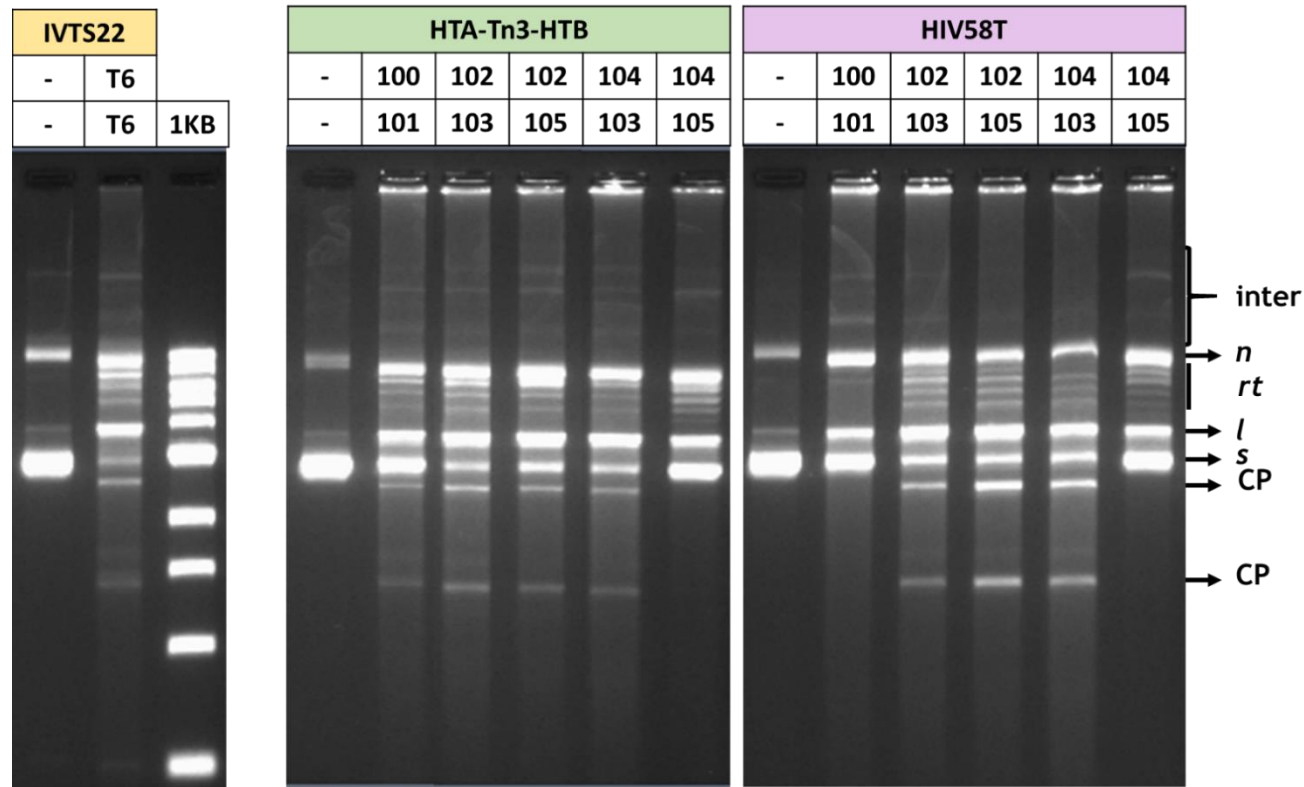


Figure 5.14: *In vitro* recombination activities of HIV TALERs on HTA-Tn3-HTB and HIV58T substrate plasmids (Uncut). The reactions here contained a mixture of two proteins in 1:1 ratio. Final protein concentration in the reaction was kept constant at 400 nM and DNA concentration at 25 $\mu\text{g}/\text{ml}$. The activity of TALER6 on IVTS22 serves as a control. The least amount of substrate depletion is observed with the mixture of TALER104 and TALER105. HIV-TALERS show more activity on pHIV58T than on pHTA-Tn3-HTB. The abbreviations are as follows: “s” (supercoiled substrate), “n” (nicked substrate), “l” (linear substrate), “CP” (double-site cleavage products), “rt” (recombination topoisomers), “inter” (intermolecular recombination products)

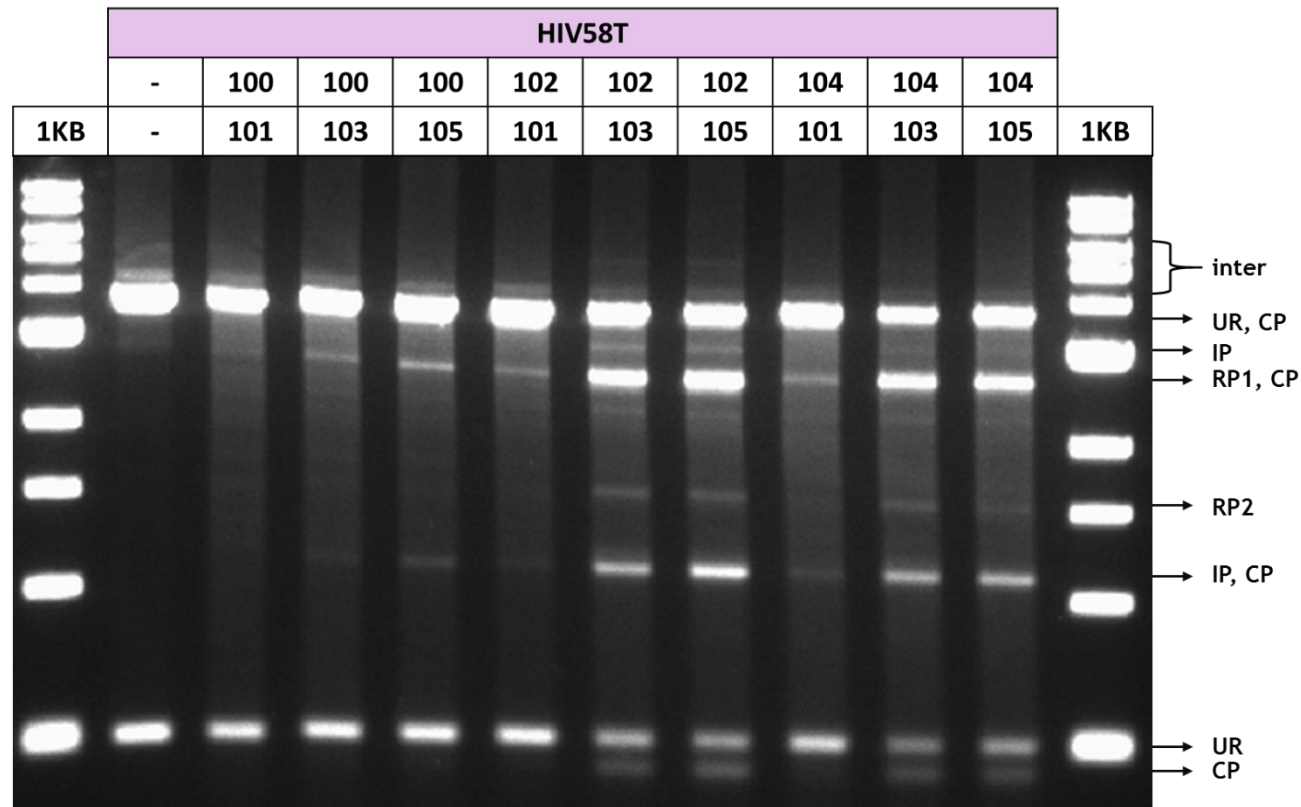


Figure 5.15: *In vitro* recombination activities of TALER100, TALER101 and HIV TALERs on HIV58T substrate plasmid (NruI-digested). The reactions here contained a mixture of two proteins in 1:1 ratio. Final protein concentration in the reaction was kept constant at 400 nM and DNA concentration at 25 μ g/ml. TALERs with Tn3 NM catalytic domain did not show significant activity on HIV58T even when used along with HIV TALERs. Reaction product sizes are as indicated in Figure 2.10A. The smaller resolution product, RP2, is definitive of recombination activity.

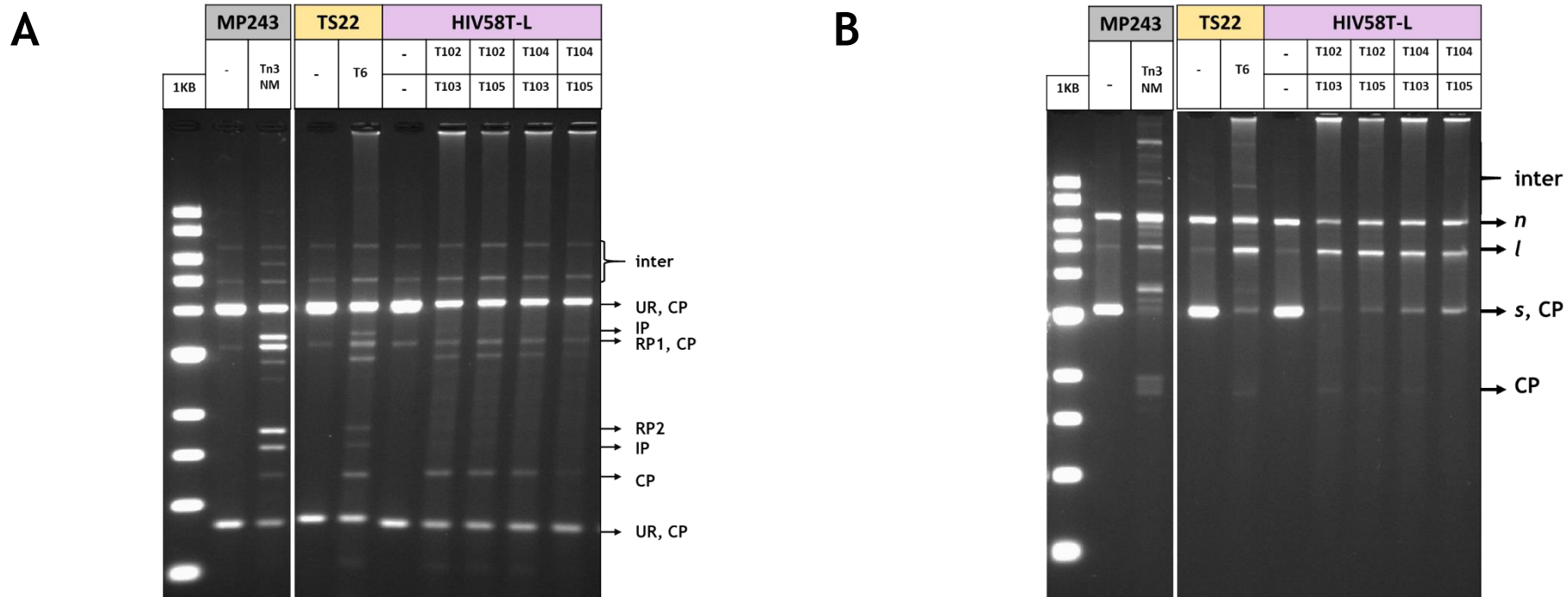


Figure 5.16: *In vitro* recombination activities of HIV TALERs on *in vivo* HIV58T(-L) substrate plasmid. The substrate plasmid tested, pJHFAB3 (HIV58T-L) is an *in vivo* recombination substrate plasmid with the 58-bp CR_TATA_target site flanking the *galK* gene in direct repeat. The activities of Tn3 NM resolvase on its *in vivo* recombination substrate plasmid (pMP243) and of TALER6 on its *in vivo* recombination substrate, TS22, serve as control reactions. Except for the control reactions, the reactions here contained a mixture of two proteins in 1:1 ratio. Final protein concentration in the reaction was kept constant at 400 nM and DNA concentration at 20 $\mu\text{g}/\text{ml}$. The blank reaction contained no protein and is marked as “-”. **A.** NruI-digested reaction products. Recombination product sizes are as indicated in Figure 2.10A. The smaller resolution product, RP2, is definitive of recombination activity. **B.** Uncut recombination reaction products. Abbreviations are as described in Fig. 5.11.

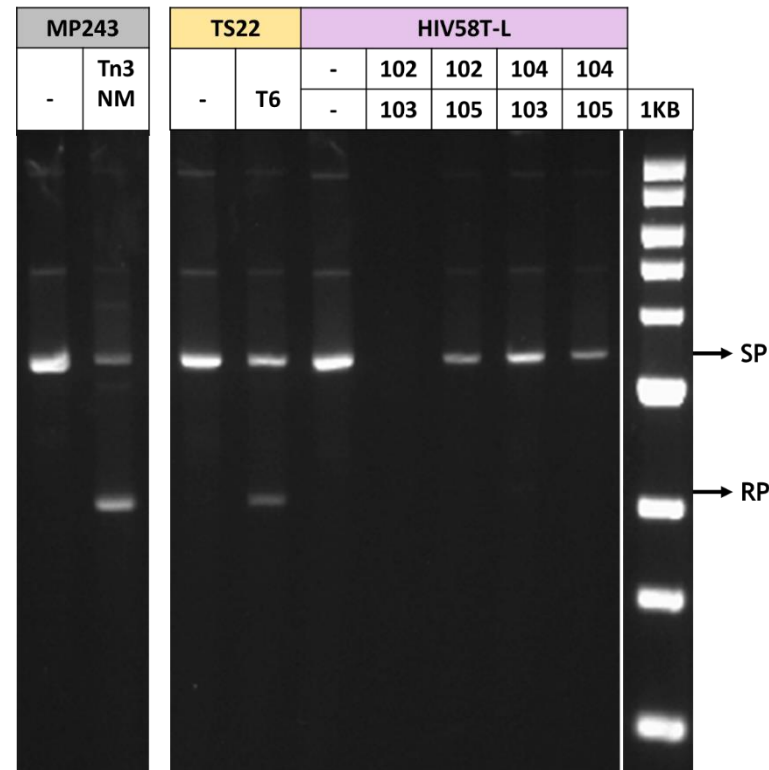


Figure 5.17: Analysis of DNA prepared from the transformation of ethanol-precipitated *in vitro* recombination reactions in Figure 5.16. Ethanol-precipitated reaction products were used to transform *E. coli* cells. After expression, the cells were inoculated into L-broth and grown overnight. Prepared DNA was run on agarose gels. The abbreviations are as follows: “SP” (Substrate plasmid), “RP” (Resolution Product). Tn3 NM resolvase and TALER6 show significant substrate depletion and resolution products on agarose gel. The activities of HIV TALERS on pJHFAB3 (HIV58T-L) is evidenced by significant substrate depletion but no observable resolution products. The combination TALER102/TALER103 did not yield any viable transformants harbouring observable DNA on agarose gel.

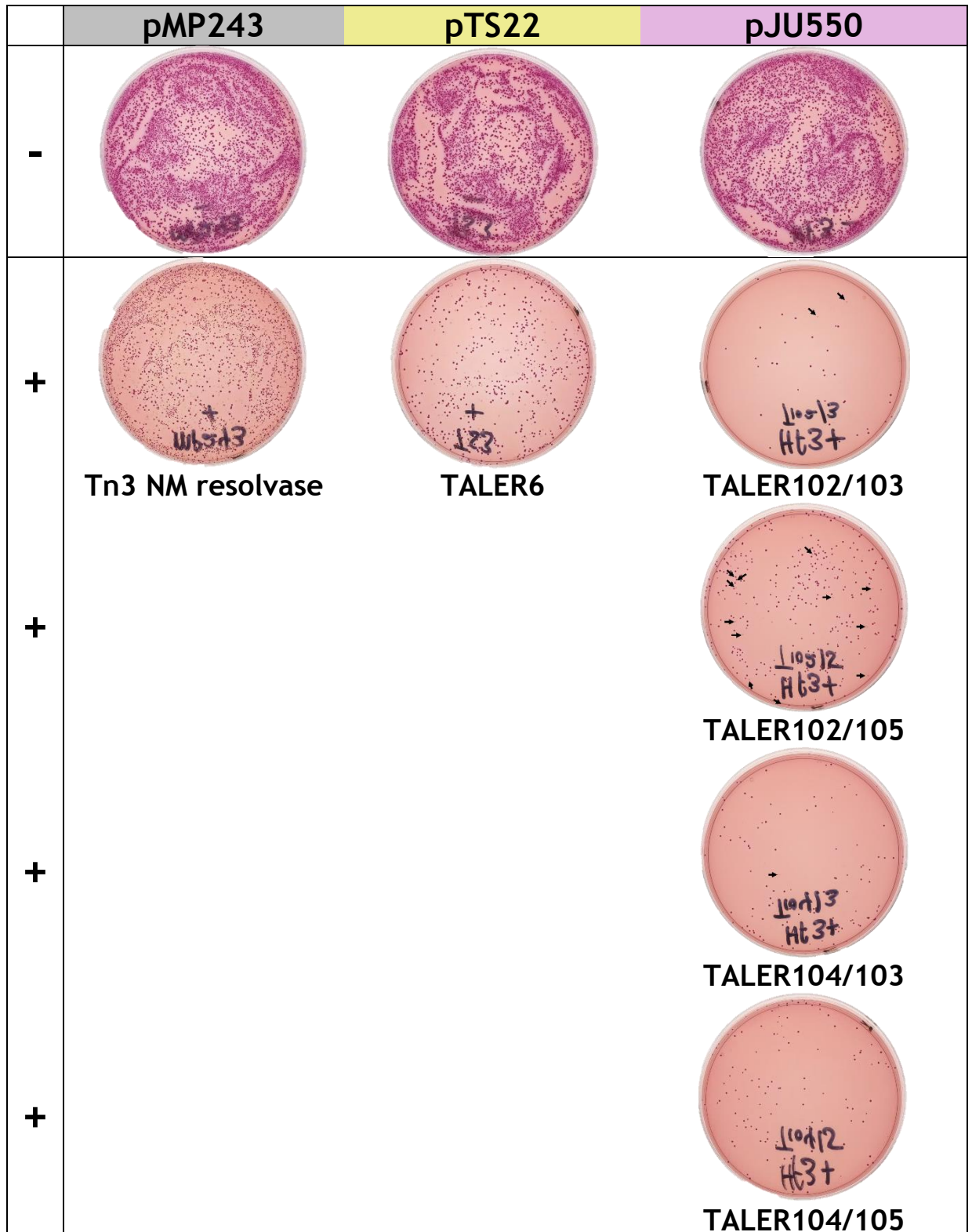


Figure 5.18: MacConkey agar plates showing cells transformed with ethanol-precipitated *in vitro* recombination reactions from Figure 5.16: “-” indicates the blank reactions without protein added and “+” with protein(s) added. Protein names are indicated below each plate. White colonies are indicated with black arrows (except on Tn3 NM res/pMP243 and TALER6/TS22). With the activities of Tn3 NM resolvase on pMP243 and TALER6 on TS22, white colonies are present at an almost 1:1 ratio with the red colonies. Significant reduction in the number of colonies can be observed with the activity of the HIV substrates on pJU550 (HIV58T-L).

5.3 Discussion

5.3.1 TALERs demonstrate significant programmable targeting capacity

In analysing the activities of TALER100, TALER101 and the HIV TALERs, it is clear that TALE proteins provide a high level of programmability for retargeting resolvase catalytic activity. The $\Delta 148/+63$ TALE architecture for TALER design from the TALE1297-based TALERs described in Chapter 4 was portable to the synthetic HIV-TALEA and HIV-TALEB contexts. The conservation of N- and C-terminal sequences of TALE proteins mean that further design iterations of TALERs that yield better activity than reported in this work can be easily cloned. It has been further demonstrated that unlike targeting with zinc finger proteins, there are not many restrictions to the sequences that can be targeted with TALEs (Section 5.2.4). Asymmetric sites as presented in the CR_TATA_target sequence and the HTA-Tn3-HTB T-site were also easily recombined by combining two different proteins as shown (Fig. 5.11 and Fig. 5.15). Non-specific targeting was not observed with any of the proteins tested; this level of orthogonality is desired for genome editing. The *in vivo* complementation ability of ZFRs on asymmetric sites has been previously reported by Proudfoot *et al.* (2011) where the co-expression of two ZFRs was required for excision at such target sites.

The three types of TALE DBDs used in this work seem to show different activities in TALER contexts, yielding recombination products to different degrees. However, the question of the effect of TALE DBD affinity for its target sequence on TALER activity has not been probed here (Fig. 5.9). This might provide unique insight for the future design of TALERs (Section 6.2).

5.3.2 HIV TALERs specifically target CR_TATA_target site

The HIV TALERs designed by combining catalytic domains from ZR012 and ZR014 with the left and right ($\Delta 148/+63$) HIV-TALE DBDs showed significant cleavage activity on the substrate plasmid carrying the 58-bp consensus CR_TATA_target sequence. This plasmid represents a mock-HIV construct with the 'LTRs' in direct

repeat and the reported activity here corresponds to site-specific excision of the proviral DNA. However, the challenging issue here is the reduced capacity of the proteins to religate the cut DNA. The actual observed levels of resolution products were quite low. It is also important to note that NM-TALER variants do not demonstrate activity on pHIV58T, and the observed activities are a function of the mutant catalytic domains. The accumulation of double-site cleavage products at levels higher than previously observed indicated that the TATA-activating mutations in the catalytic domains were significant and specific for CR_TATA_target site recombination. The highest level of activity seem to stem from the combination of TALER102 and TALER105. This combination has the left TALE DBD harbouring the 012 cat. dom. and the right TALE DBD harbouring the 045 cat. dom. An explanation for this cannot be easily made as it could result from an interplay of the TALE target sequence, RVD affinity and the catalytic domain activity.

While *in vivo* analysis within the ZFR context shows that the catalytic domain of ZR045 yields more resolution products with the full HIV-ZLR target sequence compared to that of ZR012, *in vitro* activity characterization within the TALER context seems to present a slightly different result. ZR045 and ZR012 appear to function optimally when combined in a reaction. This difference in activity could result from many factors including protein solubility in the *in vitro* test system. This must be taken into consideration in the interpretation of these results. Other catalytic domain mutant variants identified in this work (such as from ZR113 and ZR052) can also be tested in HIV-TALER contexts to check if better resolution activity can be observed (Section 3.3).

Although the results presented here suggest that TALER activity lead to the accumulation of cleavage products, the therapeutic application of these proteins for proviral DNA excision might still prove effective. The TALERS designed in this work would potentially inactivate the HIV provirus, much like a TALEN would. TALERS could however bring in a higher level of precision due to the requirement for two sites flanking the provirus for the formation of a synaptic complex.

Chapter Six: Conclusions

The primary aim of this project was to engineer a Chimaeric Recombinase that has the ability to specifically excise the HIV-1 proviral DNA. To a large extent, this objective has been met. A Tn3-based resolvase catalytic domain that targets the central recombination target sequence of CR_TATA_target site and catalyses the excision of what is between them has been designed and characterised. The resolution product from the activity of a ZFR variant of this protein has been sequenced and verified as the expected product.

After designing the HIV-targeting catalytic domain, an approach to engineer an optimal architecture for TALERs was embarked on. The analysis of several N-terminal truncation and linker length variants showed that the $\Delta 148$ truncation remains the best cut-off point for Tn3-based TALER design. It was also observed that the length of the linker between the recombinase catalytic domain and TALE DBD did not significantly alter activity. However, the distribution of recombination products from TALER activity on its target substrates was spread across several modes of DNA rearrangement- integration, inversion and resolution. A major feature of observed TALER activity was the accumulation of cleavage products. The HIV TATA-targeting TALERs designed and analysed in this work specifically target the 58-bp CR_TATA_target consensus sequence of several HIV-1 clades with precision and specific activity. However, these TALERs retained the phenomenon of cleavage product accumulation which seems to be characteristic of TALER activity.

6.1 Biotechnological implications of programmable CRs

It has been demonstrated in this work that the Tn3 resolvase catalytic domains can be quickly engineered to target difficult non-cognate sequences of clinical importance. The huge disparity between the sequences at the central 16-bp of CR_TATA_target site and Tn3 *res* site I (Fig. 3.3), and the abilities of designed proteins (like ZR045 and ZR052), to catalyse specific excision and yield precise resolution products on the target HIV-Z substrates is proof of this. The multiplicity of small serine recombinase catalytic domains available to work with and the diversity of their target sequences open up the potential targeting applications even further (Olorunniji *et al.*, 2016).

Not only has it been demonstrated that the nucleotide at position -3 (on any half of the Tn3 site I-based crossover site) plays a very important role in the catalysis of recombination, ZFRs capable of recognizing sequences with altered position -3 nucleotides were designed as well. Future work could take the identity of the nucleotide at this important position as a basis for the selection of the starting serine recombinases for reprogramming.

In contrast to the engineering of tyrosine recombinases, the mutations in these chimaeric recombinases are centred on the catalytic domain. If need be, optimised variants of the catalytic domain (of ZR045) can be designed in a single round of site-directed mutagenesis. With a protein like Brec1, which took 145 rounds to generate, its optimization or retargeting might require many more mutational rounds or a fresh start with the Cre recombinase (Karpinski *et al.*, 2016).

Although the critical recombination property of target site religation after cleavage has not been fully shown with TALERs, it is important to note that even the marginal level of precise religation that these proteins demonstrate is an improvement on current nuclease-based technologies. It can also be claimed that the current state of TALER design presents a more specific programmable 'nuclease' with inherent sequence recognition in its catalytic domain. This could maximize the fidelity of TALE targeting and the sequence-specificity of the reprogrammed resolvase catalytic domain for increased on-target activity. While the *in vitro* based assay used here to probe TALER activity presents a high concentration of substrate plasmids to the TALER protein, the target sites within a single cellular genome for a genomic excision application will actually be very limited. This implies that actual *in vivo* genomic TALER activities might be better than currently demonstrated *in vitro*. A moderate level of TALER-mediated religation in addition to host repair mechanisms might yield higher precision in genomic editing than demonstrated by TALENs, ZFNs and the CRISPR-Cas9 systems.

It is therefore important to test the current proteins in mock-HIV eukaryotic systems to analyse their actual targeting properties.

6.2 Improving TALER activity: Structural Considerations

A number of questions remain to be answered on the design and optimisation of TALERs for potential therapeutic genome editing. The optimal length of the TALE DBD target sequence needs to be determined. It is possible that a reduction in the number of TALE repeats binding at the T-sites might stabilize intramolecular recombination and/or drive resolution to completion. According to Boch *et al.* (2009), TALEs with at least 10.5 repeats in the CRD region were able to drive strong reporter gene activation. Substrate plasmids with T-sites having TALE DBD target sequences ranging from 11 bp to 17 bp could be designed to characterize the effect of reduced TALE DBD target sequence length on TALER.

Another key factor that was not probed in this work is the implication of the affinity of the TALE DBD for its target site to recombination activity. Three different TALE DBDs, harbouring a range of RVDs of different binding affinities for their target nucleotides, were used in this work. The activities of the TALERs designed from them (TALER6, TALER100 and TALER101), although having the same $\Delta 148/+63$ architecture, differed in some respects. A systematic approach to probe this by testing TALER6 on substrate plasmids bearing T-sites with lower RVD/DNA affinity might yield significant insight for TALER design. Three altered target sequences can be generated by replacing 1, 2 and 4 nucleotides in the TALE-DBD target sequence with nucleotides which the RVDs show reduced binding affinity for based on data by Moore *et al.* (2014) (Fig. 1.4). T-sites generated from these target sequences can be used to test the effect of loose TALER binding on recombination activity.

6.3 Improving TALER activity: Technical Considerations

The critical challenge in the optimization of TALERs for this work was the throughput limitation imposed due to the use of *in vitro* systems. Each design iteration for testing TALER activity had to be cloned, expressed and purified before testing. Low solubility and/or cellular toxicity of some proteins also

increased these design challenges. It is possible that the full range of activities of the TALERs designed in this work have not been fully assessed as the *in vitro* systems used in this work might not completely mimic intracellular systems. While TALE proteins are bacterial virulence factors, they are designed to function in plant cells and are foreign to *E. coli*. Their non-specific binding activity while searching for their target sites (Section 1.3.1.2) might reduce cellular viability. For future TALER activity optimization, the use of simple eukaryotic systems such as yeast cells might improve throughput and activity. Site-directed mutagenesis, and/or rational design approaches can then be used to create new TALER variants. The ease and robustness of engineering using living systems, as evidenced by the design of the HIV-targeting catalytic domain variants, could lead to the selection of more active TALER architectures.

Another potential approach for rapid and high-throughput TALER optimization is the use of droplet microfluidic systems. Droplet microfluidic technologies aim to reduce the cost and time of synthetic biology design by replacing reaction tubes with millions of nano-sized droplets. The use of cell-free protein expression systems can allow the testing of multiple iterations of TALER design in a single experiment. This has the potential to increase design scale, avoid cellular toxicity issues and explore the full sequence space for the design of optimal TALERs for genome editing.

6.4 Outlook: Genome editing for proviral DNA excision

Concerning CR-based genome editing for proviral DNA excision, the next step for this work will be the testing of the HIV TALER constructs in mock-HIV infected human cells such as HELA or HEK293 cells. The targeted substrate ‘provirus’ could be transiently introduced into these cells using a plasmid that localises in the nucleus or stably integrated into the genome using transposons, homologous recombination or viral delivery (Kim and Eberwine, 2010). The protein (fused with a nuclear localisation tag), encoded as mRNA or on a plasmid DNA, could then be delivered to the cells *in vitro* using lipofection, electroporation or cationic polymers. The substrate proviral DNA could be designed such that proviral excision leads to the reconstitution or activation of a split fluorescence reporter gene such

as split GFP (Feng *et al.*, 2017) or the loss of reporter function like luciferase activity.

Several challenges remain to be tackled for the use of CRs for proviral DNA eradication from latent reservoirs. In fact, these apply to all other genome editing approaches for this cause. Latently resting memory CD4⁺ T cells which are found in lymphoid organs and in the blood represent the primary reservoirs for latent HIV infection. However, other long-lived cells such as hematopoietic stem/progenitor cells (HSPCs), monocyte-derived cells and dendritic cells have been implicated in viral latency (Sacha and Ndlovu, 2016). The locations of these reservoirs are also diverse; ranging from the brain and bone marrow to the genital tract. These present unique targeting challenges for genome editing. It is yet unclear how many of the reservoirs need to be cleared and to what extent to prevent relapse in the patients. The persistence of latently infected cells in sanctuary sites such as tissue reservoirs has limited curative approaches to HIV in the past, and even small numbers of reservoirs are predicted to be sufficient to rekindle viral replication (Hutter, 2016; Henrich *et al.*, 2014). On a positive note, due to the direct recognition and targeting of the proviral DNA by genome editing tools such as CRs, genome editing approaches could potentially compliment other ART and reservoir-clearing strategies for improved outcomes (Sebastian and Collins, 2014).

While *ex vivo* targeting might be sufficient in other therapeutic genomic editing applications, the clearing of persistent latent reservoirs might require *in vivo* gene therapy. Significant safety and efficacy studies will need to be carried out to determine the best delivery tools and techniques for this. Off-target genotoxicity and immunogenicity of the CRs and the chosen delivery vehicles (e.g. viral vectors) must be assayed while dosage, administration routes, duration of gene editing and quality control are key issues for consideration (Shim *et al.*, 2017). There are also pressing ethical concerns about genome editing that need to be addressed.

Bibliography

- Agapakis, C. (2014). Designing Synthetic Biology. *ACS Synthetic Biology*, 3(3), 121-128.
- Ahlenstiel, C. L., Suzuki, K., Marks, K., Symonds, G. P., & Kelleher, A. D. (2015). Controlling HIV-1: Non-coding RNA gene therapy approaches to a functional cure. *Frontiers in Immunology*, 6, 474.
- Akopian, A., He, J., Boocock, M. R., & Stark, W. M. (2003). Chimeric recombinases with designed DNA sequence recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 100(15), 8688-91.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002). *Molecular biology of the cell*. New York: Garland Science.
- Alhadrami, H. (2017). Biosensors: Classifications, medical applications and future prospective. *Biotechnology and Applied Biochemistry*.
- Amado, R. G., Mitsuyasu, R. T., Rosenblatt, J. D., Ngok, F. K., Bakker, A., Cole, S., ... Symonds, G. P. (2004). Anti-Human Immunodeficiency Virus Hematopoietic Progenitor Cell-Delivered Ribozyme in a Phase I Study: Myeloid and Lymphoid Reconstitution in Human Immunodeficiency Virus Type-1-Infected Patients. *Human Gene Therapy*, 15(3), 251-262.
- Arnold, P. H., Blake, D. G., Grindley, N. D. F., Boocock, M. R., & Stark, W. M. (1999). Mutants of Tn3 resolvase which do not require accessory binding sites for recombination activity. *The EMBO Journal*, 18(5), 1407-1414.
- Arts, E. and Hazuda, D. (2012). HIV-1 Antiretroviral Drug Therapy. *Cold Spring Harbor Perspectives in Medicine*, 2(4).
- Avila, P., Ackroyd, A. and Halford, S. (1990). DNA binding by mutants of Tn21 resolvase with DNA recognition functions from Tn3 resolvase. *Journal of Molecular Biology*, 216(3), 645-655.
- Barré-Sinoussi, F., Laura Ross, A., & Delfraissy, J.-F. (2013). Past, present and future: 30 years of HIV research. *Nature Publishing Group*, 11, 877-883.
- Bednarz, A., Boocock, M. and Sherratt, D. (1990). Determinants of correct *res* site alignment in site-specific recombination by Tn3 resolvase. *Genes & Development*, 4(12b), 2366-2375.
- Bernstein, D. L., Le Lay, J. E., Ruano, E. G., & Kaestner, K. H. (2015). TALE-mediated epigenetic suppression of CDKN2A increases replication in human fibroblasts. *Journal of Clinical Investigation*, 125(5), 1998-2006.
- Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A. and Bonas, U. (2009). Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science*, 326(5959), 1509-1512.
- Bogdanove, A., Schornack, S. and Lahaye, T. (2010). TAL effectors: finding plant genes for disease and defense. *Current Opinion in Plant Biology*, 13(4), 394-401.
- Boocock, M.R., Zhu, X., and Grindley, N.D. (1995) Catalytic residues of $\gamma\delta$ resolvase act in cis. *EMBO J.* 14, 5129-5140.
- Buchholz, F. and Stewart, A. (2001). Alteration of Cre recombinase site specificity by substrate-linked protein evolution. *Nature Biotechnology*, 19(11), 1047-1052.

- Burke, M. E., Arnold, P. H., He, J., Wenwieser, S. V. C. T., Rowland, S. -J., Boocock, M. R., & Stark, W. M. (2004). Activating mutations of Tn3 resolvase marking interfaces important in recombination catalysis and its regulation. *Molecular Microbiology*, 51(4), 937-948.
- Carroll, D. (2011). Genome Engineering With Zinc-Finger Nucleases. *Genetics*, 188(4), 773-782.
- Cermak, T., Doyle, E. L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., ... Voytas, D. F. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Research*, 39(12), e82-e82.
- Chaikind, B., Bessen, J. L., Thompson, D. B., Hu, J. H., & Liu, D. R. (2016). A programmable Cas9-serine recombinase fusion protein that operates on DNA sequences in mammalian cells. *Nucleic Acids Research*, 44(20), 9758-9770.
- Chambers, S., Prior, S., Barstow, D. and Minton, N. (1988). The pMTL nic- cloning vectors. I. Improved pUC polylinker regions to facilitate the use of sonicated DNA for nucleotide sequencing. *Gene*, 68(1), 139-149.
- Chandrasegaran, S. and Carroll, D. (2016). Origins of Programmable Nucleases for Genome Engineering. *Journal of Molecular Biology*, 428(5), 963-989.
- Christian, M., Qi, Y., Zhang, Y., & Voytas, D. F. (2013). Targeted Mutagenesis of *Arabidopsis thaliana* Using Engineered TAL Effector Nucleases. *Genes|Genomes|Genetics*, 3(10), 1697-1705.
- Chun, T., Justement, J. S., Moir, S., Hallahan, C. W., Maenza, J., Mullins, J. I., ... Fauci, A. S. (2007). Decay of the HIV Reservoir in Patients Receiving Antiretroviral Therapy for Extended Periods: Implications for Eradication of Virus. *The Journal of Infectious Diseases*, 195(12), 1762-1764.
- Coates, C., Kaminski, J., Summers, J., Segal, D., Miller, A. and Kolb, A. (2005). Site-directed genome modification: derivatives of DNA-modifying enzymes as targeting tools. *Trends in Biotechnology*, 23(8), 407-419.
- Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., ... Zhang, F. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, 339(6121), 819-823.
- Cong, L., Zhou, R., Kuo, Y., Cunniff, M. and Zhang, F. (2012). Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature Communications*, 3(1).
- Cooper, G. (2017). *DNA Rearrangements*. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates; 2000.
- Cox, D., Platt, R. and Zhang, F. (2015). Therapeutic genome editing: prospects and challenges. *Nature Medicine*, 21(2), 121-131.
- Cuculis, L., Abil, Z., Zhao, H. and Schroeder, C. (2015). Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nature Communications*, 6, p.7277.
- Dampier, W., Nonnemacher, M. R., Sullivan, N. T., Jacobson, J. M., & Wigdahl, B. (2014). HIV Excision Utilizing CRISPR/Cas9 Technology: Attacking the Proviral Quasispecies in Reservoirs to Achieve a Cure. *MOJ Immunology*, 1(4).

- DeKaveler, R., Choi, V., Moehle, E., Paschon, D., Hockemeyer, D., & Meijnsing, S. *et al.* (2010). Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. *Genome Research*, 20(8), 1133-1142.
- Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J., Shi, Y. and Yan, N. (2012). Structural Basis for Sequence-Specific Recognition of DNA by TAL Effectors. *Science*, 335(6069), 720-723.
- DiGiusto, D. L., Cannon, P. M., Holmes, M. C., Li, L., Rao, A., Wang, J., ... Zaia, J. A. (2016). Preclinical development and qualification of ZFN-mediated CCR5 disruption in human hematopoietic stem/progenitor cells. *Molecular Therapy. Methods & Clinical Development*, 3, 16067.
- Ding, Q., Lee, Y., Schaefer, E., Peters, D., Veres, A., & Kim, K. *et al.* (2013). A TALEN Genome-Editing System for Generating Human Stem Cell-Based Disease Models. *Cell Stem Cell*, 12(2), 238-251.
- Ding, Y., Li, H., Chen, L. and Xie, K. (2016). Recent Advances in Genome Editing Using CRISPR/Cas9. *Frontiers in Plant Science*, 7.
- Dong, S., Lin, J., Held, N., Clem, R., Passarelli, A., & Franz, A. (2015). Heritable CRISPR/Cas9-Mediated Genome Editing in the Yellow Fever Mosquito, *Aedes aegypti*. *PLOS ONE*, 10(3), e0122353.
- Doudna, J. and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213), 1258096-1258096.
- Ebina, H., Kanemura, Y., Misawa, N., Sakuma, T., Kobayashi, T., Yamamoto, T., & Koyanagi, Y. (2015). A High Excision Potential of TALENs for Integrated DNA of HIV-Based Lentiviral Vector. *PLOS ONE*, 10(3), e0120047.
- Ebina, H., Misawa, N., Kanemura, Y., & Koyanagi, Y. (2013). Harnessing the CRISPR/Cas9 system to disrupt latent HIV-1 provirus. *Scientific Reports*, 3(1), 2510.
- Elrod-Erickson, M., Rould, M., Nekludova, L. and Pabo, C. (1996). Zif268 protein-DNA complex refined at 1.6Å: a model system for understanding zinc finger-DNA interactions. *Structure*, 4(10), 1171-1180.
- Felgner, S., Kocijancic, D., Frahm, M., & Weiss, S. (2016). Bacteria in Cancer Therapy: Renaissance of an Old Concept. *International Journal of Microbiology*, 2016, 1-14.
- Feng, S., Sekine, S., Pessino, V., Li, H., Leonetti, M. and Huang, B. (2017). Improved split fluorescent proteins for endogenous protein labeling. *Nature Communications*, 8(1).
- French, C., Mora, K., Joshi, N., Elfick, A., Haseloff, J., & Ajioka, J. (2011). Synthetic biology and the art of biosensor design. In: Institute of Medicine (US) Forum on Microbial Threats. The Science and Applications of Synthetic and Systems Biology: Workshop Summary. Washington (DC): *National Academies Press (US)*; 2011. A5.
- Fu, Y., Rocha, P. P., Luo, V. M., Raviram, R., Deng, Y., Mazzoni, E. O., & Skok, J. A. (2016). CRISPR-dCas9 and sgRNA scaffolds enable dual-colour live imaging of satellite sequences and repeat-enriched individual loci. *Nature Communications*, 7, 11707.
- Gaj, T., Mercer, A. C., Gersbach, C. A., Gordley, R. M., & Barbas, C. F. (2011). Structure-guided reprogramming of serine recombinase DNA sequence specificity.

Proceedings of the National Academy of Sciences of the United States of America, 108(2), 498-503.

Gaj, T., Mercer, A. C., Sirk, S. J., Smith, H. L., Barbas, C. F., & III. (2013). A comprehensive approach to zinc-finger recombinase customization enables genomic targeting in human cells. *Nucleic Acids Research*, 41(6), 3937-46.

Gaj, T., Sirk, S. J., Tingle, R. D., Mercer, A. C., Wallen, M. C., Barbas, C. F., & III. (2014). Enhancing the specificity of recombinase-mediated genome engineering through dimer interface redesign. *Journal of the American Chemical Society*, 136(13), 5047-56.

Gao, H., Wu, X., Chai, J. and Han, Z. (2012). Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Research*, 22(12), 1716-1720.

German Advisory Committee Blood (2016). Human Immunodeficiency Virus (HIV). *Transfusion Medicine and Hemotherapy*, 43(3), 203-222.

Gersbach, C., Gaj, T., Gordley, R. and Barbas, C. (2010). Directed evolution of recombinase specificity by split gene reassembly. *Nucleic Acids Research*, 38(12), 4198-4206.

Gordley, R., Smith, J., Gräslund, T. and Barbas, C. (2007). Evolution of Programmable Zinc Finger-recombinases with Activity in Human Cells. *Journal of Molecular Biology*, 367(3), 802-813.

Grindley, N., Whiteson, K., & Rice, P. (2006). Mechanisms of Site-Specific Recombination. *Annual Review Of Biochemistry*, 75(1), 567-605.

Guilinger, J. P., Pattanayak, V., Reyon, D., Tsai, S. Q., Sander, J. D., Joung, J. K., & Liu, D. R. (2014). Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. *Nature Methods*, 11(4), 429-35.

Hauber, I., Hofmann-Sieber, H., Chemnitz, J., Dubrau, D., Chusainow, J., & Stucka, R. *et al.* (2013). Highly Significant Antiviral Activity of HIV-1 LTR-Specific Tre-Recombinase in Humanized Mice. *Plos Pathogens*, 9(9), e1003587.

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends in Molecular Medicine*, 18(3), 182-92.

Henrich, T., Hanhauser, E., Marty, F., Sirignano, M., Keating, S., Lee, T., Robles, Y., Davis, B., Li, J., Heisey, A., Hill, A., Busch, M., Armand, P., Soiffer, R., Altfeld, M. and Kuritzkes, D. (2014). Antiretroviral-Free HIV-1 Remission and Viral Rebound After Allogeneic Stem Cell Transplantation. *Annals of Internal Medicine*, 161(5), 319.

Hermann, M., Maeder, M., Rector, K., Ruiz, J., Becher, B., Bürki, K., Khayter, C., Aguzzi, A., Joung, J., Buch, T. and Pelczar, P. (2012). Evaluation of OPEN Zinc Finger Nucleases for Direct Gene Targeting of the ROSA26 Locus in Mouse Embryos. *PLoS ONE*, 7(9), p.e41796.

Hilton, I. B., D'Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, 33(5), 510-517.

Hiv.lanl.gov. (2017). *HIV Sequence Database: Nomenclature Overview*. [online] Available at: <https://www.hiv.lanl.gov/content/sequence/HelpDocs/subtypes-more.html>.

- Holt, S. E. (2014) *Target site DNA recognition by Tn3 and Sin resolvases*. PhD thesis. University of Glasgow, Glasgow, Scotland.
- Hsu, D. C., Sereti, I., & Ananworanich, J. (2013). Serious Non-AIDS events: Immunopathogenesis and interventional strategies. *AIDS Research and Therapy*, 10(1), 29.
- Huang, Z., & Nair, M. (2017). A CRISPR/Cas9 guidance RNA screen platform for HIV provirus disruption and HIV/AIDS gene therapy in astrocytes. *Scientific Reports*, 7(1), 5955.
- Hütter, G. (2016). Stem cell transplantation in strategies for curing HIV/AIDS. *AIDS Research and Therapy*, 13(1).
- Jankele, R. and Svoboda, P. (2014). TAL effectors: tools for DNA Targeting. *Briefings in Functional Genomics*, 13(5), 409-419.
- Jayaram, M., Rowley, P., Fan, H., Kachroo, A., Voziyanov, Y., Guga, P. and Ma, C. (2015). An Overview of Tyrosine Site-specific Recombination: From an Flp Perspective. *Microbiology Spectrum*, 3(4).
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337(6096), 816-821.
- Johnson, R.C., and Bruist, M.F. (1989) Intermediates in Hin-mediated DNA inversion: a role for Fis and the recombinational enhancer in the strand exchange reaction. *EMBO J.* 8, 1581-1590.
- Joung, J. K., & Sander, J. D. (2012). TALENs: a widely applicable technology for targeted genome editing. *Nature Reviews Molecular Cell Biology*, 14(1), 49-55.
- Juillerat, A., Beurdeley, M., Valton, J., Thomas, S., Dubois, G., Zaslavskiy, M., Mikolajczak, J., Bietz, F., Silva, G., Duclert, A., Daboussi, F. and Duchateau, P. (2014). Exploring the transcription activator-like effectors scaffold versatility to expand the toolbox of designer nucleases. *BMC Molecular Biology*, 15(1), p.13.
- Kaminski, R., Bella, R., Yin, C., Otte, J., Ferrante, P., Gendelman, H. E., ... Khalili, K. (2016). Excision of HIV-1 DNA by gene editing: a proof-of-concept *in vivo* study. *Gene Therapy*, 23(8 & 9), 690-695.
- Kamtekar, S., Ho, R. S., Cocco, M. J., Li, W., Wenwieser, S. V. C. T., Boocock, M. R., ... Steitz, T. A. (2006). Implications of structures of synaptic tetramers of gamma delta resolvase for the mechanism of recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 103(28), 10642-7.
<https://doi.org/10.1073/pnas.0604062103>
- Kaplan, T., Friedman, N. and Margalit, H. (2005). Ab Initio Prediction of Transcription Factor Targets Using Structural Knowledge. *PLoS Computational Biology*, 1(1), p.e1.
- Karpinski, J., Hauber, I., Chemnitz, J., Schäfer, C., Paszkowski-Rogacz, M., & Chakraborty, D. *et al.* (2016). Directed evolution of a recombinase that excises the provirus of most HIV-1 primary isolates with high specificity. *Nature Biotechnology*, 34(4), 401-409.

- Karvelis, T., Gasiunas, G., Young, J., Bigelyte, G., Silanskas, A., Cigan, M., & Siksnys, V. (2015). Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. *Genome Biology*, 16, 253.
- Khalil, A. and Collins, J. (2010). Synthetic biology: applications come of age. *Nature Reviews Genetics*, 11(5), 367-379.
- Kim, H., & Kim, J.-S. (2014). A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics*, 15(5), 321-334.
- Kim, T. and Eberwine, J. (2010). Mammalian cell transfection: the present and the future. *Analytical and Bioanalytical Chemistry*, 397(8), 3173-3178.
- Kleinstiver, B., Wang, L., Wolfs, J., Kolaczyk, T., McDowell, B., Wang, X., Schild-Poulter, C., Bogdanove, A. and Edgell, D. (2014). The I-TevI Nuclease and Linker Domains Contribute to the Specificity of Monomeric TALENs. *Genes|Genomes|Genetics*, 4(6), 1155-1165.
- Kubik, G., & Summerer, D. (2016). TALEored Epigenetics: A DNA-Binding Scaffold for Programmable Epigenome Editing and Analysis. *ChemBioChem*, 17(11), 975-980.
- Krebs FC, Hogan TH, Quiterio S, Gartner S, Wigdahl B. (2002). Lentiviral LTR-directed expression, sequence variation, and disease pathogenesis. *Los Alamos National Laboratory HIV Sequence: Compendium*.
- Kusano, H., Onodera, H., Kihira, M., Aoki, H., Matsuzaki, H. and Shimada, H. (2016). A simple Gateway-assisted construction system of TALEN genes for plant genome editing. *Scientific Reports*, 6(1).
- Laity, J., Dyson, H. and Wright, P. (2000). DNA-induced α -helix capping in conserved linker sequences is a determinant of binding affinity in Cys2-His2 zinc fingers. *Journal of Molecular Biology*, 295(4), 719-727.
- Laity, J., Lee, B. and Wright, P. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Current Opinion in Structural Biology*, 11(1), 39-46.
- Lam, K., van Bakel, H., Cote, A., van der Ven, A. and Hughes, T. (2011). Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Research*, 39(11), 4680-4690.
- Lamb, B. M., Mercer, A. C., & Barbas, C. F. (2013). Directed evolution of the TALE N-terminal domain for recognition of all 5' bases. *Nucleic Acids Research*, 41(21), 9779-9785.
- Laskey, S. and Siliciano, R. (2014). A mechanistic theory to explain the efficacy of antiretroviral therapy. *Nature Reviews Microbiology*, 12(11), pp.772-780.
- Leenay, R. T., Maksimchuk, K. R., Slotkowski, R. A., Agrawal, R. N., Gomaa, A. A., Briner, A. E., ... Beisel, C. L. (2016). Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Molecular Cell*, 62(1), 137-147.
- Lei, H., Sun, J., Baldwin, E. P., Segal, D. J., & Duan, Y. (2014). Conformational Elasticity can Facilitate TALE-DNA Recognition.
- Li, L., Wu, L. and Chandrasegaran, S. (1992). Functional domains in Fok I restriction endonuclease. *Proceedings of the National Academy of Sciences*, 89(10), 4275-4279

- Li, K., Pang, J., Cheng, H., Liu, W.-P., Di, J.-M., Xiao, H.-J., ... Gao, X. (2015). Manipulation of prostate cancer metastasis by locus-specific modification of the CRMP4 promoter region using chimeric TALE DNA methyltransferase and demethylase. *Oncotarget*, 6(12), 10030-10044.
- Li, T., Huang, S., Jiang, W. Z., Wright, D., Spalding, M. H., Weeks, D. P., & Yang, B. (2011). TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Research*, 39(1), 359-372.
- Li, W. (2005). Structure of a Synaptic Resolvase Tetramer Covalently Linked to Two Cleaved DNAs. *Science*, 309(5738), 1210-1215.
- Liang, C., Wainberg, M., Das, A., & Berkhout, B. (2016). CRISPR/Cas9: a double-edged sword when used to combat HIV infection. *Retrovirology*, 13(1).
- Liang, P., Xu, Y., Zhang, X., Ding, C., Huang, R., Zhang, Z., ... Huang, J. (2015). CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes. *Protein & Cell*, 6(5), 363-372.
- Liu, J. and Stormo, G. (2008). Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics*, 24(17), 1850-1857.
- Liu, J., Gaj, T., Patterson, J. T., Sirk, S. J., & Barbas III, C. F. (2014). Cell-Penetrating Peptide-Mediated Delivery of TALEN Proteins via Bioconjugation for Genome Engineering. *PLoS ONE*, 9(1), e85755.
- Ma, H., Marti-Gutierrez, N., Park, S.-W., Wu, J., Lee, Y., Suzuki, K., ... Mitalipov, S. (2017). Correction of a pathogenic gene mutation in human embryos. *Nature*, 548(7668), 413-419.
- Maeder, M., Thibodeau-Beganny, S., Sander, J., Voytas, D. and Joung, J. (2009). Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nature Protocols*, 4(10), 1471-1501.
- Maeder, M., Thibodeau-Beganny, S., Osiaik, A., Wright, D., Anthony, R., Eichinger, M., Jiang, T., Foley, J., Winfrey, R., Townsend, J., Unger-Wallace, E., Sander, J., Müller-Lerch, F., Fu, F., Pearlberg, J., Göbel, C., Dassie, J., Pruett-Miller, S., Porteus, M., Sgroi, D., Iafrate, A., Dobbs, D., McCray, P., Cathomen, T., Voytas, D. and Joung, J. (2008). Rapid "Open-Source" Engineering of Customized Zinc-Finger Nucleases for Highly Efficient Gene Modification. *Molecular Cell*, 31(2), 294-301.
- Maervoet, V. and Briers, Y. (2016). Synthetic biology of modular proteins. *Bioengineered*, 8(3), 196-202.
- Mak, A., Bradley, P., Cernadas, R., Bogdanove, A. and Stoddard, B. (2012). The Crystal Structure of TAL Effector PthXo1 Bound to Its DNA Target. *Science*, 335(6069), 716-719.
- Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., ... Church, G. M. (2013). RNA-Guided Human Genome Engineering via Cas9. *Science*, 339(6121), 823-826.
- Meckler, J. F., Bhakta, M. S., Kim, M.-S., Ovadia, R., Habrian, C. H., Zykovich, A., ... Baldwin, E. P. (2013). Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Research*, 41(7), 4118-4128.
- Meinke, G., Bohm, A., Hauber, J., Pisabarro, M. and Buchholz, F. (2016). Cre Recombinase and Other Tyrosine Recombinases. *Chemical Reviews*, 116(20), 12785-12820.

- Meinke, G., Karpinski, J., Buchholz, F., & Bohm, A. (2017). Crystal structure of an engineered, HIV-specific recombinase for removal of integrated proviral DNA. *Nucleic Acids Research*, 45(16), 9726-9740.
- Mercer, A. C., Gaj, T., Fuller, R. P., & Barbas, C. F. (2012). Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic Acids Research*, 40(21), 11163-11172.
- Miller, J. C., Tan, S., Qiao, G., Barlow, K. A., Wang, J., Xia, D. F., ... Rebar, E. J. (2011). A TALE nuclease architecture for efficient genome editing. *Nature Biotechnology*, 29(2), 143-148.
- Miller, J., McLachlan, A. and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *The EMBO journal*, 4 (6), 1609-14.
- Moore, R., Chandrabhas, A., & Bleris, L. (2014). Transcription Activator-like Effectors: A Toolkit for Synthetic Biology. *ACS Synthetic Biology*, 3(10), 708-716.
- Mouw, K., Rowland, S., Gajjar, M., Boocock, M., Stark, W. and Rice, P. (2008). Architecture of a Serine Recombinase-DNA Regulatory Complex. *Molecular Cell*, 30(2), 145-155.
- Mussolino, C., Morbitzer, R., Lütge, F., Dannemann, N., Lahaye, T., & Cathomen, T. (2011). A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Research*, 39(21), 9283-93.
- Niu, Q., Zheng, H., Zhang, L., Qin, F., Facemire, L., Zhang, G., Cao, F., Zhang, K., Huang, X., Yang, J., He, L. and Liu, C. (2015). Knockout of theadpgene related with colonization in *Bacillus nematocida* B16 using customized transcription activator-like effectors nucleases. *Microbial Biotechnology*, 8(4), 681-692
- Olorunniji, F., He, J., Wenwieser, S., Boocock, M., & Stark, W. (2008). Synapsis and catalysis by activated Tn3 resolvase mutants. *Nucleic Acids Research*, 36(22), 7181-7191.
- Olorunniji, F., Rosser, S., & Stark, W. (2016). Site-specific recombinases: molecular machines for the Genetic Revolution. *Biochemical Journal*, 473(6), 673-684.
- Olorunniji, F. and Stark, W. (2009). The catalytic residues of Tn3 resolvase. *Nucleic Acids Research*, 37(22), 7590-7602.
- Olorunniji, F., & Stark, W. (2010). Catalysis of site-specific recombination by Tn3resolvase. *Biochemical Society Transactions*, 38(2), 417-421.
- Owens, J. B., Mauro, D., Stoytchev, I., Bhakta, M. S., Kim, M.-S., Segal, D. J., & Moisyadi, S. (2013). Transcription activator like effector (TALE)-directed piggyBac transposition in human cells. *Nucleic Acids Research*, 41(19), 9197-207.
- Panda, S. K., Wefers, B., Ortiz, O., Floss, T., Schmid, B., Haass, C., ... Kühn, R. (2013). Highly Efficient Targeted Mutagenesis in Mice Using TALENs. *Genetics*, 195(3), 703-713.
- Peng, R., Lin, G., & Li, J. (2016). Potential pitfalls of CRISPR/Cas9-mediated genome editing. *The FEBS Journal*, 283(7), 1218-1231.
- Perez, E., Wang, J., Miller, J., Jouvenot, Y., Kim, K., & Liu, O. *et al.* (2008). Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature Biotechnology*, 26(7), 808-816.

- Polstein, L. R., & Gersbach, C. A. (2015). A light-inducible CRISPR-Cas9 system for control of endogenous gene activation. *Nature Chemical Biology*, 11(3), 198-200.
- Proudfoot, C., McPherson, A. L., Kolb, A. F., Stark, W. M., & III, C. B. (2011). Zinc Finger Recombinases with Adaptable DNA Sequence Specificity. *PLoS ONE*, 6(4), e19537.
- Prorocic, Marko Milenkovic (2009) *Sequence selectivity of the resolvase catalytic domain: implications for Z-resolvase design*. PhD thesis, University of Glasgow
- Prorocic, M., Wenlong, D., Olorunniji, F., Akopian, A., Schloetel, J., Hannigan, A., McPherson, A. and Stark, W. (2011). Zinc-finger recombinase activities *in vitro*. *Nucleic Acids Research*, 39(21), 9316-9328
- Ramalingam, S., Annaluru, N. and Chandrasegaran, S. (2013). A CRISPR way to engineer the human genome. *Genome Biology*, 14(2), p.107.
- Resources, A., & Products, G. (1998). HIV sequence database HIV - 1 Gene Map, 13-15.
- Rice, P. (2015). Serine Resolvases. *Microbiology Spectrum*, 3(2).
- Rice, P., & Steitz, T. (1994). Refinement of γ δ resolvase reveals a strikingly flexible molecule. *Structure*, 2(5), 371-384.
- Rowland, S., Stark, W. and Boocock, M. (2002). Sin recombinase from *Staphylococcus aureus*: synaptic complex architecture and transposon targeting. *Molecular Microbiology*, 44(3), 607-619.
- Rowland, S., Boocock, M., McPherson, A., Mouw, K., Rice, P. and Stark, W. (2009). Regulatory mutations in Sin recombinase support a structure-based model of the synaptosome. *Molecular Microbiology*, 74(2), 282-298.
- Sacha, J. and Ndhlovu, L. (2016). Strategies to target non-T-cell HIV reservoirs. *Current Opinion in HIV and AIDS*, 11(4), 376-382.
- Sakuma, T., Ochiai, H., Kaneko, T., Mashimo, T., Tokumasu, D., Sakane, Y., ... Yamamoto, T. (2013). Repeating pattern of non-RVD variations in DNA-binding modules enhances TALEN activity. *Scientific Reports*, 3(1), 3379.
- Sakuma, T. and Woltjen, K. (2014). Nuclease-mediated genome editing: At the front-line of functional genomics technology. *Development, Growth & Differentiation*, 56(1), 2-13.
- Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual* (2nd Ed.). 10.51-10.67.
- Sander, J., Cade, L., Khayter, C., Reyon, D., Peterson, R., Joung, J. and Yeh, J. (2011). Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nature Biotechnology*, 29(8), 697-698.
- Sander, J., Reyon, D., Maeder, M., Foley, J., Thibodeau-Beganny, S., Li, X., Regan, M., Dahlborg, E., Goodwin, M., Fu, F., Voytas, D., Joung, J. and Dobbs, D. (2010). Predicting success of oligomerized pool engineering (OPEN) for zinc finger target site sequences. *BMC Bioinformatics*, 11(1), p.543.
- Sander, J., Maeder, M., Reyon, D., Voytas, D., Joung, J. and Dobbs, D. (2010). ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool. *Nucleic Acids Research*, 38(Web Server), W462-W468

- Sanderson, M., Freemont, P., Rice, P., Goldman, A., Hatfull, G., Grindley, N., & Steitz, T. (1990). The crystal structure of the catalytic domain of the site-specific recombination enzyme $\gamma\delta$ resolvase at 2.7 Å resolution. *Cell*, 63(6), 1323-1329.
- Sanjana, N., Cong, L., Zhou, Y., Cunniff, M., Feng, G. and Zhang, F. (2012). A transcription activator-like effector toolbox for genome engineering. *Nature Protocols*, 7(1), 171-192.
- Sarkar, I., Hauber, I., Hauber, J. and Buchholz, F. (2007). HIV-1 Proviral DNA Excision Using an Evolved Recombinase. *Science*, 316(5833), 1912-1915.
- Sarkis, G., Murley, L., Leschziner, A., Boocock, M., Stark, W. and Grindley, N. (2001). A Model for the $\gamma\delta$ Resolvase Synaptic Complex. *Molecular Cell*, 8(3), 623-631.
- Schneider, F., Schwikardi, M., Muskhelishvili, G. and Dröge, P. (2000). A DNA-binding domain swap converts the invertase gin into a resolvase. *Journal of Molecular Biology*, 295(4), 767-775.
- Schreiber, T., Sorgatz, A., List, F., Blüher, D., Thieme, S., Wilmanns, M. and Bonas, U. (2015). Refined Requirements for Protein Regions Important for Activity of the TALE AvrBs3. *PLOS ONE*, 10(3), e0120214
- Sebastian, N. and Collins, K. (2014). Targeting HIV latency: resting memory T cells, hematopoietic progenitor cells and future directions. *Expert Review of Anti-infective Therapy*, 12(10), 1187-1201.
- Shalem, O., Sanjana, N. E., & Zhang, F. (2015). High-throughput functional genomics using CRISPR-Cas9. *Nature Reviews Genetics*, 16(5), 299-311.
- Shapiro, J. (1979). Molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. *Proceedings Of The National Academy Of Sciences*, 76(4), 1933-1937.
- Shim, G., Kim, D., Park, G., Jin, H., Suh, S. and Oh, Y. (2017). Therapeutic gene editing: delivery and regulatory perspectives. *Acta Pharmacologica Sinica*, 38(6), 738-753.
- Siliciano, R. and Greene, W. (2011). HIV Latency. *Cold Spring Harbor Perspectives in Medicine*, 1(1), a007096-a007096.
- Sin, M. L. Y., Mach, K. E., Wong, P. K., & Liao, J. C. (2014). Advances and challenges in biosensor-based diagnosis of infectious diseases. *Expert Review of Molecular Diagnostics*, 14(2), 225-44.
- Sirk, S., Gaj, T., Jonsson, A., Mercer, A. and Barbas, C. (2014). Expanding the zinc-finger recombinase repertoire: directed evolution and mutational analysis of serine recombinase specificity determinants. *Nucleic Acids Research*, 42(7), 4755-4766.
- Stark, W. (2014). The Serine Recombinases. *Microbiology Spectrum*, 2(6).
- Stark, W. M., Parker, C. N., Halford, S. E., & Boocock, M. R. (1994). Stereoselectivity of DNA catenane fusion by resolvase. *Nature*, 368(6466), 76-78.
- Stone, D., Kiem, H.-P., & Jerome, K. R. (2013). Targeted gene disruption to cure HIV. *Current Opinion in HIV and AIDS*, 8(3), 217-23.
- Studier, F. W., Rosenberg, A. H., Dunn, J. J. and Dubendorff, J. W. (1990). Use of T7 RNA polymerase to direct expression of cloned genes. *Methods. Enzymol.* 185, 60-89

- Summers, D. K. and Sherratt, D. J. (1988). Resolution of ColE1 dimers requires a DNA sequence implicated in the three-dimensional organization of Xer site. *EMBO J.* 7, 851-858
- Sung, P., & Klein, H. (2006). Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nature Reviews Molecular Cell Biology*, 7(10), 739-750.
- Szurek, B., Rossier, O., Hause, G. and Bonas, U. (2002). Type III-dependent translocation of the *Xanthomonas AvrBs3* protein into the plant cell. *Molecular Microbiology*, 46(1), 13-23.
- Tebas, P., Stein, D., Tang, W., Frank, I., Wang, S., & Lee, G. *et al.* (2014). Gene Editing of CCR5 in Autologous CD4 T Cells of Persons Infected with HIV. *New England Journal Of Medicine*, 370(10), 901-910.
- Tsuji, S., Futaki, S., & Imanishi, M. (2013). Creating a TALE protein with unbiased 5'-T binding. *Biochemical And Biophysical Research Communications*, 441(1), 262-265.
- van Duyne, G. D. (2015). Cre Recombinase. *Microbiology Spectrum*, 3(1).
- van Passel, M. W. J., Lam, C. M. C., Martins dos Santos, V. A. P., & Suárez-Diez, M. (2014). Synthetic Biology in Health and Disease. In *Synbio and Human Health*. Dordrecht: Springer Netherlands., 1-10.
- Vidic, J., Manzano, M., Chang, C. and Jaffrezic-Renault, N. (2017). Advanced biosensors for detection of pathogens related to livestock and poultry. *Veterinary Research*, 48(1).
- Vieira, J. and Messing, J. (1982). The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene*, 19(3), 259-268.
- Wang, C. X., & Cannon, P. M. (2016). The clinical applications of genome editing in HIV. *Blood*, 127(21), 2546-52.
- Wang, G. P., Levine, B. L., Binder, G. K., Berry, C. C., Malani, N., McGarrity, G., ... Bushman, F. D. (2009). Analysis of Lentiviral Vector Integration in HIV+ Study Subjects Receiving Autologous Infusions of Gene Modified CD4+ T Cells. *Molecular Therapy*, 17(5), 844-850.
- Wang, H., La Russa, M., & Qi, L. S. (2016). CRISPR/Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry*, 85(1), 227-264.
- Wang, Z., Pan, Q., Gendron, P., Zhu, W., Guo, F., Cen, S., ... Liang, C. (2016). CRISPR/Cas9-Derived Mutations Both Inhibit HIV-1 Replication and Accelerate Viral Escape. *Cell Reports*, 15(3), 481-489.
- Wolfe, S., Nekludova, L. and Pabo, C. (2000). DNA Recognition by Cys2His2Zinc Finger Proteins. *Annual Review of Biophysics and Biomolecular Structure*, 29(1), 183-212.
- Wang, H., La Russa, M. and Qi, L. (2016). CRISPR/Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry*, 85(1), 227-264.
- Wong, J. K., Hezareh, M., Günthard, H. F., Havlir, D. V, Ignacio, C. C., Spina, C. A., & Richman, D. D. (1997). Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science (New York, N.Y.)*, 278(5341), 1291-5.

World Health Organization. (2017). *WHO - a global health guardian*. [online] Available at: <http://www.who.int/publications/10-year-review/health-guardian/en/index3.html>.

Xu, L., Yang, H., Gao, Y., Chen, Z., Xie, L., Liu, Y., ... Deng, H. (2017). CRISPR/Cas9-Mediated CCR5 Ablation in Human Hematopoietic Stem/Progenitor Cells Confers HIV-1 Resistance *In vivo*. *Molecular Therapy*, 25(8), 1782-1789.

Yang, W. (2010). Nucleases: diversity of structure, function and mechanism. *Quarterly Reviews of Biophysics*, 44(01), 1-93.

Yang, F., Liu, C., Chen, D., Tu, M., Xie, H., Sun, H., ... Gu, F. (2017). CRISPR/Cas9-loxP-Mediated Gene Editing as a Novel Site-Specific Genetic Manipulation Tool. *Molecular Therapy: Nucleic Acid*, 7, 378-386.

Yang, J., Zhang, Y., Yuan, P., Zhou, Y., Cai, C., Ren, Q., ... Wei, W. (2014). Complete decoding of TAL effectors for DNA recognition 628 Complete decoding of TAL effectors for DNA recognition. *Cell Research*, 24(24), 628-631.

Yang, W., & Steitz, T. (1995). Crystal structure of the site-specific recombinase $\gamma\delta$ resolvase complexed with a 34 by cleavage site. *Cell*, 82(2), 193-207.

Yao, Y., Nashun, B., Zhou, T., Qin, L., Qin, L., Zhao, S., ... Chen, X. (2012). Generation of CD34⁺ Cells from CCR5-Disrupted Human Embryonic and Induced Pluripotent Stem Cells. *Human Gene Therapy*, 23(2), 238-242.

Yin, C., Zhang, T., Qu, X., Zhang, Y., Putatunda, R., Xiao, X., ... Hu, W. (2017). *In vivo* Excision of HIV-1 Provirus by saCas9 and Multiplex Single-Guide RNAs in Animal Models. *Molecular Therapy*, 25(5), 1168-1186.

Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G. and Arlotta, P. (2011). Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature Biotechnology*, 29(2), 149-153.

Zheng, J. H., Nguyen, V. H., Jiang, S.-N., Park, S.-H., Tan, W., Hong, S. H., ... Min, J.-J. (2017). Two-step enhanced cancer immunotherapy with engineered *Salmonella typhimurium* secreting heterologous flagellin. *Science Translational Medicine*, 9(376).

Appendix

The sequences of some of the plasmids used in this study are provided here. Key features are annotated. Plasmid information and descriptions are provided in Section 2.6 and Section 2.10.

1. pJUM004

ZR004 M13 REV ColEI origin AmpR M13 FWD

```

1 GAATTCTAGAAAAAATTTTGGTTAACTTTAAGAAGGAGATATACATATGGCCATTTTTGGTTATGCACGCGTCTCAACCA 80
81 GCCAGCAGTCCCTCGATATTCAGATCAGAGCGCTCAAAGATGCAGGGGTAAGCTAACCCGATCTTTACCGACAAGCT 160
161 TCCGGCTCGAGCACCATCGGGGAGGGCTGGATTTGCTTCGAATGAAGGTGAAGGAAGGTGACGTCATTTCTGGTGAAGAA 240
241 GCTCGACCGGTTAGGCCGCGACACCGCCGACATGCTCCAAGTATAAAAGAGTTTGTGCTCAGGGGTGATAGCGGTTCCGT 320
321 TTATCGATGACGGGATCAGTACCGACTCCTACATCGGGCTGATGGTTGTCACCATCCTGTCTGCAGTGGCACAGGCTGAG 400
401 CGTTGAGGATCTTAGAGCGCACGAATGAGGGCAGACAGGCAGCAAAGCTGAAAGGAGTCAAATTTGGTGCAGCGGCTAC 480
481 CGTGGACAGGACTAGTGAACGTCGGTATGCTTTGCCGTTGAATCCTGTGACCGTCTGTTCTCGAGATCAGACGAAGTGA 560
561 CCCGTCACATCCGTATCCACACCGGTCAGAAAACCGTCCAGTCCGCTATATGCATGAGGAACTTCTCCAGATCTGACCAC 640
641 CTGACCACCCACATCCGTACGCACACTGGCGAAAACCGTTCGATGTCGATATCTGCGGTGTTAAATTCGCGCGCTCTGA 720
721 TGAACGTAACGTCACACCAAAATCCACTGCGTCAGAAAAGATTCGAGCTCATGAGGGTACCCTAGAGCTTGAGTATTC 800
801 TATAGTGTACCTAAATAGCTTGGCGTAATCATGGTCAAGCTGTTCCCTGTGTGAAATTTGTTATCCCGCTCACAAATCCA 880
881 CACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCTTAATGAGTGAAGTAACTCAATTAATTCGCTTGGC 960
961 CTCACCTCCCGCTTTCCAGTCGGGAAACCTGTGCTGCGCAGTGCATTAATGAATCGGCCAACCGCGGGGAGAGCGGTT 1040
1041 TGCGTATTGGGCGCTCTTCCGCTTCCCTCGTCACTGACTCGCTGCGCTCGGTCGTTCCGGTGCAGCGGATCAGCT 1120
1121 CACTCAAAGGCGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGAATTAATTCATGTTTGAC 1200
1201 AGCTTATCATaGATTAGCTTTAATGCGGTAGTTTATACAGTAAATTTGCTAACGCAGTCAGGCACCGTGTATGAAATCT 1280
1281 AACAAATCGCTCATCGTCACTCCCTCGCACCTCACCTCGGATGATGAGGATAGGCTATGAGGATAGGCTTGTATGCGCGG 1360
1361 CCTCTTGGCGGATCGACGCGAGGCTGGATGGCCTTCCCATTATGATTCTTCTCGCTTCCGGCGGCATCGGGATGCCCGC 1440
1441 GTTGCAGGCCATGCTGTCCAGGAGGTAGATGACGACCATCAGGGACAGCTTCAAGGATCGCTCGCGGCTTTACCAGCC 1520
1521 TAACCTCGATCACTGGACCGCTGATCGTCACGGCAGTTTATCCGCGCTCGGCGAGCACATGGAACGGGTTGGCATGGAT 1600
1601 GTAGGCGCCGCTATACTTGTCTGCCCTCCCGGCTTGGCTGCGGTCGAGTGCATGAGCGGGCCACCTCGACCTGAATGGA 1680
1681 AGCCGGCGGCACCTCGCTAACGGATTACCACCTCCAAGAATGGAGCCAATCAATTTCTGCGGAGAACTGTGAATGCCGA 1760
1761 AACCAACCTTGGCAGAACATATCCATCGCGTCCGCCATCTCCAGCAGCCGACGCGGGCGCATCTCGGGCAGCGTTGGGT 1840
1841 CCTGGCCACGGGTGCGCATGATCGTGTCTGCTGAGGACCCGGCTAGGCTGGCGGGTGTGCTTACTGGTTAGCAG 1920
1921 AATGAATCACCGATACGCGAGCGAAGCTGAAGCGACTGCTGCTCAAAAACGCTGCGACCTGAGCAACACATGAATGGT 2000
2001 CTTCCGTTTCCGTGTTTCGTAAGTCTGGAACCGGGAAGTCAGCGCCCTGCACCATTATGTTCCGGATCTGCATCGCAG 2080
2081 GATGCTGTGGCTACCCGTGTTGAACACCTACATCTGTATTAACGAAGCGCTGGCATGACCTGAGTGAATTTTCTCTGG 2160
2161 TCCCGCCGATCCATACCGCCAGTTGTTTACCCTCACAAACGTTCCAGTAAACGGGATGTTTATCATCAGTAACCCGAT 2240
2241 CGTGAGACTCTCTCTCGTTTTCATCGGTATCATTACCCTGATGAACAGAAATTTCCCTTACACGGAGGCATCAAGTAC 2320
2321 CAAACAGGAAAAAACCGCCCTTAAATGGCCGCTTTATCAGAAGCCAGACATTAACGCTTCTGGAGAACTCAACGAGC 2400
2401 TGACGCGGGATGAACAGGCGACATCTGTGAATCGCTTACGACCCAGCTGATGAGCTTTACCGCAGCTGCCTCGCGCT 2480
2481 TTCGGTGTGACGGTGAACACCTCTGACACATGACGCTCCCGGAGACGGTCAAGCTTGTCTGTAAGCGGATGCGGGAG 2560
2561 GACACAAGCCCGTCAGGGCGCTGAGCGGGTGTGGCGGGTGTGCGGGCGCAGCCATGACCCGATCAGCTAGCAGTACGG 2640
2641 GAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATatATGCGGTGTGAAATACCGCA 2720
2721 CAGATGCGTAAGGAGAAAAACCGCATCAGGCGCTCTTCCGCTTCCCTCGCTCACTGACTCGCTGCGCTCGGTCTGCTGG 2800
2801 TGCGCGGAGCGGTATCAGCTCACTCAAAGGCGTAATACGTTGATCCACAGAATCAGGGGATAACGCAGGAAAGAACATG 2880
2881 TGAGCAAAGGCGCAGCAAAGGCCAGAACCGTAAAAAGGCCCGTGTGCTGGGCTTTTCCATAGGCTCCGCCCCCTGA 2960
2961 CGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAAACCCGACAGGACTATAAAAGATACAGGCGTTCCTCCCTG 3040
3041 GAAGCTCCCTCGTGCCTCTCCGTTCGACCCCTGCGGCTTACCGGATACCTGTCCGCTTTCTCCCTTCGGGAAGCGTG 3120
3121 GCGCTTTCTCATAGCTACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCCGCTCCAAGCTGGGCTGTGTGCACGAAC 3200
3201 CCCGTTTACGCCCAGCCGCTGCGCTTATCCGCTAACCTTCTGAGTCCAACCCGGTAAGACACGACTTATCCGCCAC 3280
3281 TGCGAGCCCACTGGTAAACAGGATTAAGCAGAGCGAGGTATGTAAGCGGTGCTACAGAGTCTTGAAGTGGTGGCCCTAAC 3360
3361 TACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAGAAAGAGTTGGTAGCTC 3440
3441 TTGATCCGGCAAAACAAACCCAGCTGGTAGCGGTGTTTGTGTTTGTGCAAGCAGCAGATACCGCGGAAAAAAGAGAT 3520
3521 CTCAAGAAGATCCTTTGATCTTTTTCAGGGGTTGACGCTGAGCTGGAAGCAACACTCAGTTAAGGATTTTGGTCA 3600
3601 AGATTATCAAAAAGGATCTTACCTAGATCCTTTTAAATTAATAAATGAAGTTTTTAAATCAATCTAAAGTATATATGAGTA 3680
3681 AACTTGGTCTGACAGTTA CCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTTCGTTTATCCATAGTTG 3760
3761 CTTGACTCCCGCTCGTGTAGATAAATACGATACGGGAGGGCTTACCATCTGGCCCCAGTGTGCAATGATACCGCGAGAC 3840
3841 CCAGCTCACCGGCTCCAGTTTATCAGCAATAAACGCCGGAAGGGCGAGCCAGAGTCTCCGCAACTTT 3920
3921 ATCCGCTCCATCCAGTCTATTAATTTGTTGCCGGAAGCTAGAGTAAGTAGTTCCGCAAGTAAAGTTTGGCGCAACGTTG 4000
4001 TTGCCATTGCTACAGGCATCGTGGTGTACGCTCGTCTGTTGGTATGGCTTCATTCAGCTCCGGTTCACCAACGATCAAG 4080
4081 CGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGTTCCGATCGTTGTCAGAGTAAGTTGGC 4160
4161 CGCAGTGTATCACTCATGGTTATGGCAGCACTGCATAATCTCTTACTGTCTATGCCATCCGTAAGATGCTTTTCTGTGA 4240
4241 CTGGTGAGTACTCAACCAAGTCAATTCGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTCCCGCGGCTCAACACGGGAT 4320
4321 AATACCGGCCACATAGCAGAACTTTAAAAGTGTCAI CATTGAAAACGTTCTTCGGGGCGAAAACCTCTCAAGGATCTT 4400
4401 ACCGCTGTGAGATCCAGTTCGATGTAACCCACTCGTGCACCAACTGATCTTCAGCATCTTTTACTTTACCAGCGTTT 4480
4481 CTGGGTGAGCAAAAACAGGAAGGCAAAATCCGCAAAAAGGGAATAAGGGCGACCGGAAATGTTGAATACTCATACT 4560
4561 TTTCTTTTCAATATATTAATGAAGCATTATCAGGGTTATTTGCTCATGAGCGGATACATATTTGAATGTATTTGAAAAA 4640
4641 TAAACAAATAGGGGTTCCGCGCACATTTCCCGAAAAGTGCACCTGACGCTAAGAAACCATTATATCATGACATTA 4720
4721 CCTATAAAAAATAGGCGTATCACGAGGCCCTTTCTGCTCGCGGTTTCCGGTGTGACGGTGAAGAACTCTGACACATGCAG 4800
4801 CTCCGGAGACGGTACAGCTTGTCTGTAAGCGGATCGCGGAGCACAAAGCCGTCAGGGCGCTCAGCGGGTGTGG 4880
4881 CGGCTGTGCGGGCTGGCTTAACTATAGCGCATCAGAGCAGATTGCTACTGAGAGTGCACCATatATGCGGTGTGAATACC 4960
4961 GCACAGATGCGTAAGGAGAAAAATACCGCATCAGGCGAAATTTGTAACGTTAATATTTTGTAAAATTCGCGTTAAAATAT 5040
5041 TGTAAATCAGCTCATTTTTTAAACCAATAGGCCGAAATCGGCAAAATCCCTTATAAATCAAAGAAATAGACCGAGATAG 5120
5121 GTTGAAGTGTGTTCCAGTTTGAACAAGAGTCCACTATTAAGAAGCTGGACTCCAACGTCAAAGGGCAAAAACCGCT 5200
5201 ATCAGGGCGATGGCCCACTACGTGAACCATCAACCAAAATCAAGTTTTTTTTCGCGTTCGAGGTGCGCTAAAGCT 5280
5281 AACCTAAAGGGAGCCCCGATTTAGAGCTTGACGGGAAAGCCGGCAACGTTGGCGAGAAAGGAAGGAAGAAAGCGAA 5360
5361 AGGAGCGGGCGTAGGGCGCTGGCAAGTGTAGCGGTACGCTGCGCGTAACCACCACCCCGCGCTTAAATGCGCCG 5440
5441 TACAGGGCGGCTCCATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCTCTTCGCTATTACGCC 5520

```

5521 AGCTGGCGAAAGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAAACGCCAGGGTTTTCCAGTCACGACGT TGTAAAACG 5600
5601 ACGGCCAGTGAATTGTAATACGACTCACTATAGGGC 5636

2. pJUM204H

ZR004

T7

ColE1 origin

KanR

1 GCATGCAAGGAGATGGCGCCCAACAGTCCCCGGCCACGGGGCTGCCACCATACCACGCCGAAACAAGCGCTCATGAG 80
81 CCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCCATATAGGCGCCAGCAACCCGACCTGTGGCGCCGGTGA 160
161 TGCCGGCCACGATGCGTCCGGCTAGAGGATCGAGATCTCGATCCCGCGAAAT TAATACGACTCACTATAGGGGAATTGT 240
241 GAGCGGATAACAATTCCCTCTAGAAAATAATTTGTTTAACTTAAAGAGGAGATATACATATGCGCCATTTTTGGTTATG 320
321 CACGGCTCTCAACCAGCCAGCAGTCCCTCGATATTCAGATCAGAGCGCTCAAAGATGCAGGGGTTAAAAGCTAACCGCATC 400
401 TTACCACAAAGCTTCCGGCTCGAGCACCATCGGGAGGGCTGGATTGCTTCGAATGAAGGTGAAGGAAGGTGACGT 480
481 CATTCTGGTGAAGAAGCTCGACCGTTAGGCCGCGACACCGCCGACATGCTCCAAGTATAAAAAGAGTTTGTATGCTCAGG 560
561 GTGTAGCGGTTCCGTTTATCGATGACGGGATCAGTACCAGCTCCTACATCGGGCTGATGGTTGCACCACTCCTGTCTGCA 640
641 GTGGCACAGGCTGAGCGCTTGGATCTTAGAGCGCACGAATGAGGGCAGACAGGCAGCAAAGCTGAAAGGAGTCAAATT 720
721 TGGTCGACGGCGTACCGTGGACAGGACTAGTGAACGTCGATGCTTGTCCGGTTGAATCCTGTGACCGTCTGTTCTCGA 800
801 GATCAGACGAAGTACCCGTCACATCCGATCCACACCGGTCAGAAACCGTTCCAGTCCCGTATATGCATGAGGAACTTC 880
881 TCCAGATTAACCTGTGATAAACTACCGCATTAAAGCTTATCGATGATAAGCTGCAAAACATGAGAAATCTTGCAGTCTGAA 960
961 ATTCGCGCTCTGATGAACGTAACGTCACACCAAAATCCACCTGCGTCAGAAAGATTTCAGACTCACATCACCATCACC 1040
1041 ATCACTAATAAGCTCGGGTACTCTAGAGTGCATCCCGCTGCTAACAAAGCCCGAAAGGAAGGTGAGTTGGCTGCTGCC 1120
1121 ACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCTTAAACGGGCTTGGAGGGTTTTTGTGAAAGGAGGAACAT 1200
1201 ATCCGGTATCCACAGGACGGGTGTGGTCGCCATGATCGCGTAGTCGATAGTGGCTCCAAGTCCAGGACAGGACT 1280
1281 GGGCGCGGCCAAAGCGGTGGACAGTGTCCGAGAACGGGTGCGCATAGAAATGTCATCAACGCATATAGCGCTAGCAG 1360
1361 CACGCCATAGTACTGGCGATGCTGTCCGAATGGACGATATCCCGCAAGAGGCCCGGACGATACCGGCATAACCAGCCTA 1440
1441 TGCTACAGCATCCAGGTTGACGGTGGCCGAGGATGACGATGAGCGCATTTAGATTTCATACACGGTGCCTGACTGCGT 1520
1521 TAGCAATTAACCTGTGATAAACTACCGCATTAAAGCTTATCGATGATAAGCTGCAAAACATGAGAAATCTTGCAGTCTGAA 1600
1601 AGGGCCTCGTGATACGCCATTTTTATAGGTTAATGTCATGATAATAATGGTTTTCTTAGACGTCAGGTGGCACTTTTCGG 1680
1681 GAAATGTGCGCGGAACCCCTATTTGTTATTTTCTAAATACATCAAATATGATCCGCTCATGAGACAATAACCCCTG 1760
1761 ATAAATGCTTCAATAATGACCTGCAGGGGGGGGGGAAAGCCAGCTTGTGTCCAAAATCTCTGATGTTACATTTGCACA 1840
1841 AGATAAAATATATCATCATGAACAATAAAAGTGTCTGCTTACATAAAACAGTAATAACAAGGGGTGTATGAGCCATATTC 1920
1921 AACGGAAACGCTTGTCTGAGGCGCGGATTAATTCACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGAT 2000
2001 AATGTCGGCAATCAGGTGCGACAATCTATCGATTGATGGAAGCCCGATGCGCCAGAGTGTGTTCTGAAACATGGCAA 2080
2081 AGGTAGCGTTGCCAATGATGTTACAGATGAGATGGTCAGACTAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCA 2160
2161 AGCATTATCCGACTCCTGATGATGATGGTTACTCACCTGCGATCCCGGGGAAAACAGCATTCCAGGATATAGAA 2240
2241 GAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCAGTGTCTCTGCGCGGTTGCATTCGATTCCTGTTTGTAA 2320
2321 TTGTCCTTTAAACAGCGATCGCTATTTCTGCTCGCTCAGGCGCAATCAGAAATGAATAACGGTTTGGTTGATGCGAGTG 2400
2401 ATTTGATGACGAGCGTAATGGCTGGCTGTTGAACAAGTGTGAAAGAAATGCATAAGCTTTTGGCATTCTCACCAGGAT 2480
2481 TCAGTCTCACTCATGGTGATTTCTCACTTGATAACCTTATTTTACGAGGGGAAATTAATAGGTTGATTTGATGTTGG 2560
2561 ACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAAGTGCCTCGGTGAGTTTTCTCTTCATTACAGA 2640
2641 AACGGCTTTTCAAAAATATGGTATGATAATCCTGATGATAAATGTCAGTTTCATTTGATGCTCGATGAGTTTTTC 2720
2721 TAA TCAGAAATGGTTAATTTGGTTGTAACACTGGCAGAGCATTACGCTGACTTGACGGGACGGCGGCTTTGTTGAATAAAT 2800
2801 CGAATTTTGTGAGTTGAAGGATCAGATCAGCATCTTCCCGCAACGCAGACCGTTCCGTCAGGTCAGGCAAAAGTTCAA 2880
2881 AATCACCAACTGGTCCACCTACAACAAAGCTCTCATCAACCGTGGCTCCCTCACTTTCTGGCTGGATGATGGGGCGATT 2960
2961 AGGCCTGGTATGAGTCAGCAACACCTTCTCAGAGGAGACCTCAGCGCCCCCCCCCTGAGGTCAAAAGGATCTA 3040
3041 GGTGAAGATCCTTTTGTATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTTCGTTCCACTGAGCGTCAGACCCCGTAG 3120
3121 AAAAGATCAAAGGATCTTCTGAGATCCTTTTTTCTGCGCTAATCTGCTGCTGCAAAACAAAACCCAGCCGTACCA 3200
3201 GCGGTGGTTGTTTCCGGATCAAGAGTACCAACTCTTTTCCGAAGGTAACGGCTTCAGCAGAGCGCAGATACCAAA 3280
3281 TACTGTCTTCTAGTGTAGCCGTAGTTAGGCCACCACCTCAAGAACTCTGTAGACCCGCTACATACCTCGCTCTGCTAA 3360
3361 TCCTGTTACAGTGGCTGCTGCCAGTGGCGATAAGTCTGTCTTACCGGGTGGACTCAAGACGATAGTTACCGGATAAC 3440
3441 GCGCAGCGGTCGGCTGAACGGGGGTTCTGTCGACACAGCCAGCTTGGAGCGAACGACTACACCGAATGAGATACT 3520
3521 ACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCGAAAGGAGAAAGGGGACAGGTATCCGGTAAAGCGCGAGGGTCGGAA 3600
3601 CAGGAGAGCGCACGAGGAGCTTCCAGGGGAAACGCCTGGTATCTTTATAGTCTGTGCGGGTTTCGCCACCTCTGACTT 3680
3681 GAGCGTCGATTTTGTGATGCTCGTCAGGGGGCGGACGCTATGGAAAAACGCCAGCAACCGCGCCTTTTTACGGTTCTCT 3760
3761 GCGTTTTGCTGGCCTTTGCTCACATGTTCTTCTGCTGATTTCCCTGATTTCTGTGATAACCGTATACCGCTTTG 3840
3841 AGTGAGCTGATACCGCTCGCCGAGCCGAACGACCGAGCGCAGCGAGTCACTGAGCGGAAAGCGGAAGCGCCTGATG 3920
3921 CGGTATTTCTCCTTACGCATCTGTGCGGTATTTACACCCGCATATATGGTGCCTCTCAGTACAATCTGCTCTGATGCC 4000
4001 GCATAGTTAAGCCAGTATACACTCCGCTACGCTACGTGAGTGGGTCAAGGCTGCGCCCGCAGACCCGCCAACACCCGCT 4080
4081 GACGCGCCTGACGGGCTGTCTGCTCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCA 4160
4161 GAGGTTTTACCGTCATCACCAGAAACGCGGAGGCAGCTGCGGTAAGCTCATCAGCGTGGTCTGAAAGCGATTACACAGA 4240
4241 TGCTGCTGTTTATCCGGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGCTGGCTTCTGATAAAGCGGGCCATG 4320
4321 TTAAGGGCGGTTTTTCTGTTTGGTCACTGATGCCTCCGTTTAAAGGGGATTTCTGTTTACGGGGTAAATGATACCGAT 4400
4401 GAAACGAGAGGATGCTCAGATACGGGTTACTGATGAACTCCCGGTTACTGAACTGTTGAGGGTAAACAAAC 4480
4481 TGCGGTTATGGATGCGCGGGGACAGAGAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTTGTT 4560
4561 CCACAGGTTAGCCAGCAGCATCTGCGATGCGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACT 4640
4641 TTAGCAAAACCGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTGCGAGACGTTTTCGAGCAGCAGTCTGCTTCACTT 4720
4721 GCTCGCTATCGGTGATTCATTTCTGCTAACCAAGCAACCCGACGCTAGCCGGTCTCAACGACAGGACGACG 4800
4801 ATCATGCGCACCCGTTGGCCAGGACCAACGCTGCCAGCTGCCGCGCCGCTGCGGCTGCTGGAGATGGCGGACCGGATGGA 4880
4881 TATGTTCTGCCAAGGTTGGTTTGGCATTACAGTTCTCCGCAAGAATGATTGGCTCCAATCTTGGAGTGGTGAATC 4960
4961 CGTTAGCGAGGTGCCCGCGGCTTCCATTCAGGTCGAGGTGGCCCGGCTCCATGCACCGCGCAACCGCGGGGAGGAG 5040
5041 CAAGGTATAGGGCGCGCCTACAATCCATGCCAACCGTTCCATGTGCTCGCCGAGCGGCATAAATCGCCGCTGACGAT 5120
5121 AGCGGTCAGTGATCGAAGTTAGGCTGTTAAGAGCCGCGAGGACTCCTTGAAGCTGTCCTGATGGTCTCATCTACCTG 5200
5201 CCTGGACAGCATGGCTGCAACGCGGCATCCCGATGCCCGCGAAGCGAGAAGAAATCATAATGGGAAGGCCATCCAGC 5280
5281 CTCGCTCGGGCCGCTGCCCGCGATAATGGCTGCTTCTCGCGAAACGTTTGGTGGCGGGACAGTACGAAAGGCTT 5360
5361 GAGCGAGGGCTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCTGCGGCTCCAGCGAAAGCGGCTCTCGCCG 5440

5441 AAAATGACCCAGAGCGCTGCCGGCACCTGTCTACGAGTGCATGATAAAGAAGACAGTCATAAGTGGCGGACGATAGT 5520
5521 CATGCCCCGCGCCACCAGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTGAGATCCCGTGCCTAATGAG 5600
5601 TGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTCCAGTCGGGAAACCTGTCGTGCCAGTGCATTAATGA 5680
5681 ATCGGCCAACCGCGGGGAGAGGGGTTTGGCGTATTGGGACCGAGGTTGTTTTTCCACCATGAGACGGGCAACA 5760
5761 GCTGATTGCCCTTACCAGCTGGCCCTGAGAGAGTTGACAGCAAGCGGTCCACGCTGGTTTGGCCAGCAGGCGAAAATCC 5840
5841 TGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTGGGTATCGTGTATCCCACTACCGAGATATCCGCACC 5920
5921 AACCGCGAGCCCGGACTCGGTAATGGCGCGCATTTGCCGCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAA 6000
6001 CGATGCCCTCATTACGACTTTGCATGGTTTGTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCCGC 6080
6081 TGAATTTGATTGCGAGTGAATATTTATGCCAGCAGCCAGACGACGAGCGCGGAGACAGAAGTAAATGGGCGCCGCTAA 6160
6161 CAGCGCGATTTGCTGGTACCCAATGCGACAGATGCTCCACGCCAGTCGCGTACCCTTTCATGGGAGAAAATAATAC 6240
6241 TGTTGATGGGTGCTGTTGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGACGAGCTCCACAGCAATGGCATCC 6320
6321 TGGTCATCCAGCGGATGTTAATGATCAGCCCACTGACCGGTTGCGCGAGAAGATGTGACCCCGCTTACAGGCTTC 6400
6401 GACGCGCTTCCGTTCTACCATCGACACCACCGTCCAGCCAGTTGATCGGCGGAGATTTAATCGCCGCGCAAAATTT 6480
6481 GCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGGCCCGCAGTTGTTGTGCCAG 6560
6561 CGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTCCCGCTTTTCCGAGAAACGTGGCTGGCTGGTT 6640
6641 CACCACGGGGAAACGGTCTGATAAGAGACACCGGCATCTCGGACATCGTATAACGTTACTGGTTTACATTTACCA 6720
6721 CCCTGAATTGACTCTCTTCCGGCGCTATCATGCCATACCCGCAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCG 6800
6801 ACGCTCTCCCTTATGCGACTCTGCATTAGGAAGCAGCCAGTAGTAGTTGAGGCGGTTGAGCACCGCCGCCGCAAGGA 6880
6881 ATGGT 6885

3. pJU003

HIV-ZLR Z-site M13-Fwd GalK siteB-141 (F) siteA-141 (F) KanR

Tue May 15, 2018 16:59 +0100.apeMap

1 GCATCGTGGTGTACGCTCGGCAtggcggcGCTAGCAAGATCTTCTAGTCCGTGGGCGAGCGCTGCATATAAGCAGCAG 80
81 CCGCCCACGCGCGGCCGCGCATGCATCGATAGATCCAATTGCCGTGACGCAGGCATGTTTCTCAATAACGAAATTTGATA 160
161 AAATCCCGCTCTTTCATAACATTAATTTAGCCTTCTTCAGGGCTGACTGTTGCATAAAAATTCATCTGTATGCACAATA 240
241 ATGTTGTATCAACCACCATATCGGGTGAATTTGCGAAGCTCGGCTAAGCAAGAGATTAATTAATAAAGCATTGGGGC 320
321 AACGCGAATTAATTCCTGGCCGTCGTTTACACGTCGTGACTGGGAAAACCTGGCGTTACCAACTTAATCGCCTTG 400
401 CAGCACATCCCCCTTTCGCCAGGGGCAATAAGGGCTGACGCGCACTTTTATCCGCTTCTGCTCCGACCCGCTAC 480
481 GTAATTTATGGTTGGTTATGAAATGCTGGCAGAGACCAGCGAGACCTGACCGCAGAACAGGCAGCAGAGCGTTTGGCC 560
561 CGAGTCAGCGATATCCATTTTCGCCGAATCCGGAGTGTAAAGAAATGAGTCTGAAAGAAAACACAATCTCTGTTTGC 640
641 CGCATTGGGCTACCTGCCACTCACACCATTAGCGGCTGGCCGCGTGAATTTGATGGTGAACACCCGACTACAACG 720
721 ACGGTTTCGTTCTGCCCTGCGCGATTGATTATCAAAACCGTCAAGCTTGTGACACCAGCCCGCCAGCCGTTAAGCTCG 800
801 ATGGCAGCCGATTAAGAAAACAGCTCGACGAGTTTCCCTCGATGCGCCATTGTCGCACATGAAAATCATCAATGGGC 880
881 TAACTACGTTTCGTTGGCTGGTGAACATCTGCAACTCGGTAACAACAGCTTCCGCGCGCTGGACATGGTGATCAGCGG 960
961 ATGTGGCCGACGGGTGCCGGGTTAAGTCTTCCGCTTCACTGGAAGTCCGCGGTGGAAACCGTATTCGAGCAGCTTTAT 1040
1041 TTTCCGCTGGACGGCGCACAAAACCGCGCTTAACGGCTCAGGAAGCAGAAAACCAAGTTTGAAGCTGCGGGACTG 1120
1121 GGATCAGCTAATTTCCGCGCTCGGCAAGAAAGATCATGCCTTGTGATCGATTGCCGCTCACTGGGGACCAAGCAGTTT 1200
1201 CCATGCCCAAAGGTGTGGCTGTGCTCATCATCAACAGTAACTTCAAACGTAACCTGGTTGGCAGCGAATACAACCCCGT 1280
1281 CGTGAACAGTGGCAACCCGGTGGCGGTTTCTTCCAGCAGCCAGCCCTGCGTGTATGTCACATTGAAGAGTTCAACCGCT 1360
1361 TCGCCTTCGTTTCCACCCGATCGTGGCAAAACCGTGGTATCAGTGTGACAAAACCGCCAGCCGTTGAAGCTGCCA 1440
1441 GCGCGCTGGAGCAAGGCGACCTGAAACGATAGGGGAGTTGATGGCGGAGTCTCATGCCCTATGCGCGATGATTTGCA 1520
1521 ATCACCCTGCGCGCAAATGACACTCTGGTAGAATCGTCAAAGCTGTGATTGGCGACAAGGTTGGCGTACGCATGACCG 1600
1601 CGGCGGATTTGGCGGCTGTATCGTCCGCTGATCCCGAAGAGCTGGTGCCTGCCGCACGCAAGCTGTCCGTGAACAAT 1680
1681 ATGAAGCAAAAACAGGTTAATAAGAGACTTTTACGTTTGTAAACCATCACAAAGGAGCAGGACAGCTGCTGAACGAACT 1760
1761 CCGCACTGCAGGATCgatacCATATGACGTCGACCGCTGTCAGAAGCTTCTAGGTGAATTCGCGTGGGCGAGCGCTGCAT 1840
1841 ATAAGCAGCAGCCGCCCCACGCGAGCTCCCGGTACCAAtggCGGTGAACAGTGTGTTCTACTTTTGTGTTAGTCTTGATG 1920
1921 CTTCACTGATAGATACAAGGCCATAAGAACCCTCAGATCCTCCGTATTTAGCCAGTATGTTCTCTAGTGTGGTTCGTTG 2000
2001 TTTTTCGCTGAGCCATGAGAACAACCATTTGAGCTCAGCTTACCTTTGATGTCATCAAAAATTTGCTCAAACCTGG 2080
2081 TGAGCTGAATTTTTCAGTTAAAGCATCGTGTAGTGTTTTCTTAGTCCGTTACGTAGGTAGGAATCTGATGTAATGGTT 2160
2161 GTTGGTATTTTTCACCATTCATTTTTATCTGTTGTTCTCAAGTTCGGTTACGAGATCCATTTGTCTATCTAGTTCAAC 2240
2241 TTGAAAATCAACGTATCAGTCGGGCGGCTCGCTTATCAACCACCAATTTTATATTGCTGTAAGTGTTTAAATCTTTAC 2320
2321 TTATTGGTTTCAAACCCATTTGTTAAGCCTTTAAACTCATGGTAGTTATTTTTCAAGCATTAACATGAACCTTAAATTC 2400
2401 TCAAGGCTAATCTCTATATTTGCCTTGTGAGTTTCTTTTGTGTTAGTCTTTTAAATAACCCTCATAAATCCTCATAGA 2480
2481 GTATTTGTTTCAAAGACTTAAACATGTTCCAGATTAATTTTATGAATTTTTTAACTGGAAAAGATAAGGCAATATCT 2560
2561 CTTCACTAAAACCTAATTTCTAATTTTTCGTTGAGAACTTGGCAGTGTGTTGTCACCTGGAAAATCTCAAAGCCTTTAAC 2640
2641 AAAGGATTCCTGATTTCCACAGTTCCTGTCATCAGCTCTCTGTTGCTTTAGCTAATAACACCATTAAGCATTTTCCCTACT 2720
2721 GATGTTTATCATCTGAGCGTATTGGTTATAAGTGAACGATACCGTCCGTTCTTTTCTTGTAGGGTTTCAATCGTGGGGT 2800
2801 TGAGTAGTGCCACACAGCATAAAAATAGCTTGGTTTCAATGCTCCGTTAAGTCATAGCAGCTAATCGCTAGTTCATTTGCT 2880
2881 TTGAAAACAACCTAATTCAGACATACATCTCAATTTGGTCTAGGTGATTTAATCACTATAACCAATTTGAGATGGGCTAGTCA 2960
2961 ATGATAATTAAGTCTAGTCTTTTCCCTTTGGAGTTGGGGATCTGTAAGTATTTTCAAGCATTAACATGAACCTTAAATTT 3040
3041 CTGCTAGACCCTCTGTAATTTCCGCTAGACCTTTGTGTTGTTTTTTTTGTTTATATTCAAGTGGTTATAATTTATAGAATA 3120
3121 AAGAAAAGATAAAAAAAGATAAAAAAGATAGATCCAGCCCTGTGTATAAATCACTACTACTTTAGTCAGTTCCGCGATTA 3200
3201 CAAAAGGATGTGCCAAAACCGTGTGTTGCTCCTTACAAAACAGACCTTAAACCCCTAAAGGCTTAAAGTACACCCCTCGCA 3280
3281 GCTCGGGCAAATCGCTGAATATTCCTTTTGTCTCCGACCATCAGGCAGCTGAGTGCCTGCTTTTTCGTTGACATTCAGTT 3360
3361 CGCTGCGCTCACGGCTCTGGCAGTGAATGGGGTAAATGGCACTACAGGCGCCTTTTATGGATTATGCAAGGAAAACCTAC 3440
3441 CCATAATACAAGAAAAGCCCGTACGGGCTTCTCAGGGCGTTTTATGGCGGGTCTGCTATGTTGTTGCTATCTGACTTTTT 3520
3521 GCTGTTCCAGCAGTTTCTGCCCTCTGATTTCCAGCTTACCACCTCTCGGATTAATCCCGTACAGGCTATTCAGACTGGCTA 3600
3601 ATGCACCCGTAAGGCAGCGGTATCATTAACAGCCTTACCAGCTTACTGTCGCGGATCCGTCGACCTGCAGAGGGGGGG 3680
3681 GGGCGCTGAGGCTGCCCTCGTGAAGAAGGTGTTGCTGACTCATACCAGGCTGAATCGCCCCATCATCCAGCCAGAAAGT 3760
3761 GAGGGAGCCAGGTTGATGAGAGCTTTGTTGTAGGTGGACAGTTGGTGAATTTGACTTTTGGTTCGACCGGAACCGT 3840
3841 CTCGTTGTCGGGAGATGCGTGTATCTGATCCTTCAACTCAGCAAAAGTTGATTTATCAACAAAAGCCCGCTCCCGT 3920

3921 AAGTCAGCGTAATGCTCTGCCAGTGTACAACCAATTAACCAATTCTGATTAGAAAACTCATCGAGCATCAAATGAAAC 4000
 4001 TCGAATTTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAG 4080
 4081 CGAGTTCATAGGATGGCAAGATCCTGGTATCGGTCTCGCATCCGACTCGTCCAACATCAATACAACCTATTAATTTCC 4160
 4161 CCTCGTCAAAAAAAGGTTATCAAGTGAAGAAATCACCATGAGTACGACTGGAATCCGGTGAATGGCAACAGCTTATCGT 4240
 4241 ATTTCTTTCCAGACTTGTTCACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCAT 4320
 4321 TCGTGATTGCGCCTGAGCGGAGAGAAAATACGCGATCGTGTAAAAGGACAATACAAACAGGAAATCGAATGCACCCGGC 4400
 4401 GCAGGAACACTGCCAGCCATCAACAATATTTTACCTTGAATCAGGATATCTTCTAATACCTGGAATGCTGTTTTCCCG 4480
 4481 GGGATCGCAGTGTGAGTAACCATGTCATCAGGAGTACGGATAAAATGCTTGTGTTGAGAAATGGCAACAGCTTATCGT 4560
 4561 CAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTTGGCAACGCTACCTTTGCCATGTTTTCAGAAAACACTCTGGCG 4640
 4641 CATCGGGCTTCCCATACAATCGATAGATTGTGCGACCTGATTGCGCGACATTATCGCGAGCCATTTATACCCATATAAA 4720
 4721 TCAGCATCCATGTTGGAATTTAATCGCGGCCTCGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATT 4800
 4801 ACTGTTTATGTAAGCAGACAGTTTATTGTTTCATGATGATATATTTTTATCTTTGTGCAATGTAACATCAGAGATTTTGAG 4880
 4881 ACACAACGTGGCTTTCCCCCCCCCCTGCAG 4912

4. pJU203

HIV-ZLR Z-site M13-Fwd siteA-141 (F) M13-Rev ColEI origin AmpR KanR

1 ACTGCCGGCCTCTTGGGGATATCGTCCATTCGACAGCATCGCCAGTCACTATGGCGTGTCTAGCGCCATTCGCCA 80
 81 TTCAGGCTACGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTCGCTATTACGCCAGCTGGCGAAGGGGGGATGTGC 160
 161 TCGAAGCGAATTAAGTTGGTAAAGCCAGGTTTCCCGACTTCCGACTGTAATGTTGTAACACGCGCCAGT 240
 241 TATCGGATCCATATGACGTGACGCGCTGTCAGAAAGCTTCTAGAATGTACCTTAAATCGAATATCAGACACGATGTGTCT 320
 321 ATTATGCCAAAATGACGATTTAATGGACTCGAGCGAAGCCGAATtcggattatgatacaAATTGCTTAAGCCTAG 400
 401 GCGACTAGTCCGTGGGGGAGCGCTGCATATAAGCAGCAGCCGCCACCGCGCGCGCGGGTACCATGGCATGCATCGATA 480
 481 GATCCGTGCACCTGCAGGGGGGGGGGGCGCTGAGGTTGCTGCGCTGAGGAAGGTTGTTGCTGACTCATACCGACCTGAA 560
 561 TCGCCCATCATCCAGCCAGAAAGTGAAGGGAGCCAGGTTGATGAGAGCTTTGTTGATAGTGGACCAGTTGGTGATTTTTG 640
 641 AACTTTTGTCTTGGCACGGAACGGTCTGCGTGTGCGGGAAGATGCGTGATCTGATCCTTCAACTCAGCAAAGTTCGATT 720
 721 TATTCAACAAAGCCGCTCCGCTCAAGTCAGCGTAATGCTCTGCCAGTGTTACAACCAATTAACCAATTTCTGATTAGAA 800
 801 AAATCATCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTTCAATACCATATTTTGAAGAAAGCCGTTTCT 880
 881 GTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGTCTGCGACTTCCGACTCGTCCA 960
 961 ACATCAATACAACCTATTAATTTCCCTCGTCAAAAATAGGTTATCAAGTGAGAAATCACCATGAGTACGACTGAATC 1040
 1041 CCGTGAGAATGGCAAAAGCTTATGCATTTCTTTCCAGACTTGTTCACAGGCCAGCCATTACGCTCGTCATCAAAATCAC 1120
 1121 TCGCATCAACCAACCGTTATTCATTCGTGATTGCGCCTGAGCGGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTA 1200
 1201 CAAACAGGAATCGAATGCAACCCGCGCAGGAACACTGCCAGCCATCAACAATATTTTCACTGAATCAGGATATTTCTT 1280
 1281 TAATACCTGGAATGCTGTTTTCCCGGGATCGCAGTGGTGAATACCATGCATCATCAGGAGTACGGATAAAATGCTTGA 1360
 1361 TGGTCGGAAGAGGCATAAATTCGCTCAGCCAGTTAGTCTGACCATCTCATCTGTAACATCATTTGGCAACGCTACCTTTG 1440
 1441 CCATGTTTCAGAAACACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTGCGACCTGATTGCCCGACATTATC 1520
 1521 CGAGAGCCATTTATACCCATATAAATCAGCATCATGTTGGAATTTAATCGCGGCTCGAGCAAGACGTTTCCGTTGAA 1600
 1601 TATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTATTGTTTCATGATGATATATTTTTATCTTGT 1680
 1681 GC AATGTAACATCAGAGATTTTGAACACAACCTGGCTTTCCCCCCCCCCTGCAGGTCGACGGATCCATATGACGCTG 1760
 1761 AC CGCTGTCGAGAAGCTTCTAGAATGTACCTTAAATCGAATATCAGACAGGATGTGCTATTATGCCAAAATGACGATTT 1840
 1841 AATGGCTCGAGCGAAGCCGAATtcggattatgataccaattgtattgtaaacccggtGAATTCCGTGGGCGA 1920
 1921 CCGCTGCATATAAGCAGCAGCCGCCACCGGAGCTCCCGGTACCATGGCATGCATCGATAGATCTCgatcGAGGCTCG 2000
 2001 CGAGCTTGGCGTAATCATGGTTCATAGCTGTTTCTGTGTGAAATTTGTTATCCCGCTCA AATTCACACACACATACGAGCC 2080
 2081 GGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAAGTAACTCACATTAATTCGTTGCGCTCACTGCCCGCTTT 2160
 2161 CCGTCCGGAAACCTGTGCTGCCAGCTGCATTAATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATG 2240
 2241 TTCCGTTTCTCGCTCACTGACTCGCTGCGCTCGGCTCGGCTCGGCTCGGCGGAGCGGTATCAGCTCACTCAAAGCGGTAA 2320
 2321 TACGGTTATCCACAGAAATCAGGGGATAACGACGAGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAA 2400
 2401 AAGCCCGCTTGTGCGGCTTTTCCATAGGCTCCGCCCTGACGAGCATCAAAAAATCGACGCTCAAGTCAAGAGTG 2480
 2481 GCGAAAACCCGACAGGACTATAAAGATACCAGGCTTTCCCTTGAAGCTCCCTCGTGCCTTCTCTCCGACCTGC 2560
 2561 CGCTTACCGGATACCTGTCGGCTTTTCCCTTCGGGAAGCGTGGCGCTTCTCAATGCTCACGCTGTAGGTATCTCAGT 2640
 2641 TCGGTGATAGTTCGCTCCAGCTGGGCTGTGTGACAGAACCCCGCTTACGCCCGACCGCTGCGCTTATCCGGTAA 2720
 2721 CTATCGCTTGTAGTCAACCCGGTAAGACAGCTTATCGCCACTGGCAGCAGCCATGGTAACAGGATTAGCAGGCA 2800
 2801 GGTATGATAGGCGGTGCTACAGATTTTGAAGTGTGCGCTAACCTACGCTACACTAGACAGCAAGTATTTGGTATCTGC 2880
 2881 GCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGAATCCGGCAACAAACCACCGCTGGTAGCGGTGG 2960
 2961 TTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTG 3040
 3041 AC GCTCAGTGGAAACGAAAACCTCACGTTAAGGGATTTTGGTCAATGATTAATCAAAAAGGATCTTACCTAGATCCTTTTA 3120
 3121 AATTAATAAAGTAAATTTAAATCAATCTAAAAGTATATATGAGTAAACTTGGTCTGACAGTTA CCAATGCTTAATCAGTGA 3200
 3201 GGCACCTATCTCAGCGATCTGTCTATTTTCGTTCAATCCATAGTTGCGCTGACTCCCGCTCGTGTAGATAACTACGATACGGG 3280
 3281 AGGGCTTACCATCTGGCCCAAGTGTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAAC 3360
 3361 CAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCTGCAACTTTACCGCCTCCATCCAGTCTATTAATGTTGCCGGGA 3440
 3441 AGCTAGATAAGTAGTTTCGCGAGTTAATAGTTTGCAGCAAGTGTGCGCATGCTACAGGCATCGTGGTGTCAAGCTCGT 3520
 3521 CGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCATGTTGTGCAAAAAAGCG 3600
 3601 GTTAGCTCCTTCGGTCCCTCCGATCGTGTGTCAGAAAGTAAAGTGGCCGAGTGTATCACTCATGTTATGGCAGCACTGCA 3680
 3681 TAATTTCTTACTGTCTATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAAGTACTCAACCAAGTCAATTTCTGAGAAATG 3760
 3761 GTATGCGCGCAGCCAGTTGCTCTTGGCCGGCTAATATCCGCGCCACATAGCAGAACTTTAAAAGTGTCTC 3840
 3841 ATCATTGGAAAACGTTCTTCCGGGGCAAAAACCTCAAGGATCTTACCCTGTTGAGATCCAGTTCGATGTAACCCACTCG 3920
 3921 TGCACCAACTGATCTTTCAGCATCTTTTACTTTTACCAGCGTTTCTGGGTGAGCAAAAACAGGAAAGGCAAAATGCCGCAA 4000
 4001 AAAAGGGAATAAGGGCGACAGGAAATGTTGAATACTCACTCTTCTTTTCAATATTTATGAAGCATTTATCAGGGT 4080
 4081 TATTGCTCATGAGCGGATACATATTTGAATGATTTTGAAAAAATAACAAATAGGGGTTCCGCGCACATTTCCCCGAAA 4160
 4161 AGTGCCACCTG 4171

5. pJUM410 (TALER6)

T7 NM (R148) -GSGGSGTS-Δ148TALE KanR ColEI origin

```

1 GCATGCAAGGAGATGGCGCCCAACAGTCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAG 80
81 CCCGAAGTGGCGAGCCGATCTTCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCCGACCTGTGGCGCCGGTGA 160
161 TGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGT 240
241 GAGCGGATAACAATTCCCTCTAGAAAATATTTGTAACTTAAGAAGGAGATATACATATGGCCATTTTGGTTATG 320
321 CACGCGTCTCAACCAGCCAGCAGTCCCTCGATATTCAGATCAGAGCGCTCAAAGATGCAGGGGTAAGTAACCCGATC 400
401 TTTACCGACAAAGCTTCCGGCTCGAGCACCATCGGGGAGGGCTGGATTGCTTCGAATGAAGGTGAAGGAAGGTGACGT 480
481 CATCTGGTGAAGAAGCTCGACCGGTTAGGCCGCGACACCGCCGACATGATCCAATGATAAAAGAGTTTGTAGCTCAGG 560
561 GTGTAGCGGTTTCGGTTTATCGATGACGGGATCAGTACCAGACTCTACATCGGGCTGATGGTTGCCACTCTGTGCA 640
641 GTGGCACAGGCTGAACGCGGAGGATCTAGAGCGCACGAATGAGGGCCGACAGGAAGCAAAGCTGAAAGGAATCAAATT 720
721 CCGCCGACAGGCTACCGTGGACAGGGGCTCTGGCGGTTCCGGCACTAGTCCGGCAGCGAGGTCGACTTGAGGACCCTAG 800
801 GTTATTCGCAACAGCAACAGGAGAAAATCAAGCCTAAGGTCAGGAGCACCGTCCGCGCAACACCACGAGGCGCTTGTGGG 880
881 CATGGCTTCACTCATGGCATAATGTCGCGCTTTCACAGCACCTCGCGGCTTGGGAGCTTGGGAGCTGCTGCAATCAAGA 960
961 TATGATTGCGGCCCTGCCGAAGCCACGACAGGCGATTGTAGGGGTCGGTAAACAGTGGTCGGGAGCGCGAGCACTTG 1040
1041 AGGCGCTGCTGACTGTGGCGGGTGGAGCTTAGGGGGCCCTCCGCTCCAGCTCGACACCGGGCAGCTGCTGAAGATCGCGAAG 1120
1121 AGAGGGGGAGTAACAGCGGTAGAGGCGATGCACGCCTGGCGCAATGCGCTCACCGGGGCCCTTGAACCTGACCCGAGA 1200
1201 CCAGTGTGCGCAATCGCGTCACATGACGGGGGAAAGCAAGCCTGGAACCGTGAAGGTTGTGGCCGCTCTTTGTC 1280
1281 AAGACCACGGCCTTACACCGGAGCAAGTCGTGCCATTGCAAGCAATGGGGTGGCAACAGGCTCTTGGACCGGTTGAG 1360
1361 AGACTTCTCCAGTTCTCTGTCAAGCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGTGGAGGGAA 1440
1441 ACAAGCATTTGAGACTGTCCAACGGCTCCTTCCCGTGTGTGTCAAGCCACGGTGTGACCGCTGCACAAGTGGTCGCCA 1520
1521 TCGCCAGCTATGAGCGGTAAGCAGGCGTGAACAGCTTCCAGCGCTGCTGCTGTACTGTGCCAGCAATCATGAGACTG 1600
1601 ACCCCAGACCAGGTAGTCGCAATCGCGTCACATGACGGGGGAAAGCAAGCCTTGGAAACCGTGAAGGTTGTTGCCGGT 1680
1681 CCTTTGTCAAGACCACGGCCTTACACCGGAGCAAGTCGTGGCCATTGCAATAATAACGGTGGCAACAGGCTCTTGAGA 1760
1761 CGTTTCAGAGACTTCTCCAGTTCTCTGTCAAGCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGT 1840
1841 GAGGGAAACAAGCAATTGGAGACTGTCCAACGGCTCCTTCCCGTGTGTGTCAAGCCACGGTGTGACGCTTCAACAGT 1920
1921 GGTCCGCTCGCCTCGAATGGCGCGGTAAGCAGGCGTGAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATC 2000
2001 ATGGACTGACCCAGACCAGGTAGTCGCAATCGCGTCAACCGGAGGGGAAAGCAAGCCTTGGAAACCGTGAAGGTTG 2080
2081 TTGCCGCTCCTTTGTCAAGACCACGGCCTTACACCGGAGCAAGTCGTGGCCATTGCAATCCACAGCGTGGCAACAGGC 2160
2161 TCTTGAGACGGTTTCAAGACTTCTCCAGTTCTCTGTCAAGCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGT 2240
2241 CCGATGACGGAGGAAACAAGCATTGGAGACTTCCAACGGCTTCCCGTGTGTGTCAAGCCACGGTGTGCTGCAAGTTG 2320
2321 GCACAAGTGGTCGCCATCGCCTCCAATATTTGGCGGTAAGCAGGCGCTGGAACAGTACAGCGCCTGCTGCCTGTACTGTG 2400
2401 CCAGGATCATGACTGACCCAGACCAGGTAGTCGCAATCGCGTCACATGACGGGGGAAAGCAAGCCTTGGAAACCGTGC 2480
2481 AAAGTTGTTGCCGGTCTTTGTCAAGACCACGGCTTACACCGGAGCAAGTCGTGGCCATTGCAAGCAACACTCGGTGGC 2560
2561 AAACGGCTTTGAGACGGTTCAGACTTCTCCAGTTCTCTGTCAAGCCACGGGCTGACTCCCGATCAAGTTGTAGC 2640
2641 GATTGCGTCCAACGGTGGAGGAAACAAGCATTGGAGACTGTCCAACGGCTCCTTCCCGTGTGTGTCAAGCCACGGT 2720
2721 TGACGCTGCACAAGTGGTCGCCATCGCCAGCCATGATGGCGGTAAGCAGGCGCTGGAACAGTACAGCGCCTGCTGCCT 2800
2801 GTACTGTGCCAGGATCATGGACTGACACCCGAACAGTGGTGGCCATTGCTTCTAATGGGGAGGACGGCCAGCCTTGA 2880
2881 TTCATCTGTAGCCAAATGTCCAGGCCGATCCCGGTTGGTCGCTTAAACGAATGACCATGTTGCGGCTGGCATGTC 2960
2961 TTGGTGGACGACCCGCGCTCGATGCAGTCAAAAAGGCTGCTGCCTCATGCTCCCGCATTGATCAAAAAGAACCAACGGCGA 3040
3041 ATTCCCGAGAGAATTCCACCAGGAGTGGCGCATCACCATCACCATCACTGATGATCAGGTACCCTAGAGTGCATCCGGC 3120
3121 TGCTAACAAAGCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTA 3200
3201 AACGGCTTTGAGGGTTTGTGTAAGGAGGAAGTATCCGGATATCCACAGGACGGGTGGTCGCCATGATGCGC 3280
3281 GTAGTCGATAGTGGTCCAAGTAGCGAAGCGAGCAGGACTGGGCGCGGCCAAAGCGGTCGGACAGTGTCCGAGAACGG 3360
3361 GTGCGCATAGAAATGTCATCAACGCATATAGCGCTAGCAGCACGCCATAGTACTGGCGATGCTGTCGGAATGGACGATA 3440
3441 TCCCGCAAGAGGCCCGGCATACCCGATAACCAAGCCTATGCTACAGCATCCAGGGTGCAGGTTGCCGAGGATGACGAT 3520
3521 GAGCGCATTTAGATTTCAATACACGGTGCCTGACTCGGTTAGCAATTTAAGTGTGATAAAGCTTAAAGCTTAT 3600
3601 CGATGATAAGCTGTCAACATGAGAATCTTGAAGACGAAAGGGCCTCGTGATACGCCATTTTTATAGGTTAATGTCAT 3680
3681 GATAAATAAGTGTCTTAGACGCTCAGGTGGCACTTTTCGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTCTAAA 3760
3761 TACATTCAAATATGATCCGCTCATGAGACAATAAACCCTGATAAATGCTTCAATAATGACCTGCAGGGGGGGGGGAAA 3840
3841 GCCACGTTGTGTCTCAAATCTCTGATGTACATTTGCCAAGATAAAAATATATCATCATGAACAAAGGTTGCTGCT 3920
3921 TACATAAACAGTAATACAAGGGTGTATGAGCCATATTCACGGGAAACGCTCTTGCTCGAGGCCGCGATTAAATTC 4000
4001 CATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCCGGCAATCAGGTGCCACAATCTATCGATTGTATG 4080
4081 GGAAGCCCGATGCGCCAGAGTTGTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGTTCAGA 4160
4161 CTAACCTGGCTGACGGAATTTATGCTCTTCCGACCAATCAAGCATTTCCTCACTCATGATGATGATGATGATGATGAT 4240
4241 CACTGCGATCCCGGGGAAACAGCATCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGG 4320
4321 CAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTGTAATTTTCCTTTAACAGCGATCGCGTATTTTCGCTCTCGCTCAG 4400
4401 CGGCAATCAGGAATGAATAACGGTGTGGTTGATGCGAGTATTTGATGACGAGCGTAATGGCTGGCTGTGTAACAGT 4480
4481 CTTGGAAGAAATGCATAAGCTTTTGGCTTCTCAGGATTCAGTCTCACTCATGATGATGATGATGATGATGATGATGAT 4560
4561 TTTTGGACGAGGGGAAATTAATAGGTTGATTTGATGTTGGACGAGTCGGAATCGCAGACCCGATACCAGGATCTTGCCATC 4640
4641 CTATGGAACTGCCTCGGTGAGTTTCTCCTTCATACAGAAACGGCTTTTCAAAAATATGGTATTGATAATCTGATAT 4720
4721 GAATAAATGTCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATCAGAAATGGTTAATTTGGTTGTAACACTGGCAGGCA 4800
4801 TTAGCTGACTTGACGGGACGGCGGCTTTGTTGAATAAATCAAGCATTTCCTGAGTTGAAGGATCAAGGATCAGCATCAGC 4880
4881 CCGACAACGCAGACCGTTCGGTGGCAAAGCAAAAGTTCAAAATCACCACCTGGTCCACCTACAACAAAGCTCTCATCAAC 4960
4961 CGTGGCTCCCTCACTTTCTGGCTGGATGATGGGGCGATCAGGCGCTGGTATGAGTCAGCAACCCCTTCTCACAGGCGA 5040
5041 ACCTCAGCGCCCCCCCCCTGCAGGTCAAAAGGATCTAGGTGAAGATCCTTTTGTATAATTCATGACCAAAAATCCCT 5120
5121 TAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAAGGATCTTCTTGAGACTCTTTTTCGTGG 5200
5201 CGTAATCTGCTGCTTGCAAAACAAAAAACCCCGCTACCAGCGGTGGTTGTTTGGCCGATCAAGAGCTAACACTCTTT 5280
5281 TTCGAAGGTAAGTGGCTTCAGCAGAGCGCAGATACCAAAATCTGTCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTC 5360
5361 AAGAACTCTGTAGCACCCGCTACATACCTCGCTCTGCTAATCTGTTTACCAGTGGCTGCTGCGGAGTGGCGGATAAGTCTGT 5440
5441 TCTTACCGGTTGGACTCAAGACGATAGTTACCGGATAAGCCGACGCGCTCGGCTGAACGGGGGTTCTGTGCACACAGC 5520
5521 CCAGCTTGGAGCGAACGACCTACCCGAAGTGAAGATACCTACAGCGTGAAGTATGAGAAAGCGCCAGCTTCCCGAAGGG 5600
5601 AGAAAGCGGACAGGTATCCGGTAAGCGGCAGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGAAACGCCTG 5680

```

5681 **GTATCTTTATAGTCTGTGCGGGTTTCGCCACCTCTGACTTGAGCGTTCGATTTTGTGATGCTCGTCAGGGGGGCGGAGCC** 5760
5761 **TATGGAAAACGCCAGCAACCGCGGCC**TTTTTACGGTTCCTGGCCCTTTTCTGGCTTTTGTCTCATGTTCTTTCTCTGCG 5840
5841 TTATCCCTGATCTGTGGATAACCGTATTACCGCTTTTGTAGTGTGATACCGCTCGCCAGCCGAAACGACCGAGCC 5920
5921 CAGCGAGCTCAGTGGCGAGGAAGCGGAGCGCTGATCGGTTATTTCTCCTACGACTCTGCGGTATTTTACACAC 6000
6001 GCATATATGGTGCACCTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGA 6080
6081 CTGGGTCTATGGCTGCGCCCGACACCCGCCAACACCCGCTGACGGCCCTGACGGGCTTGTCTGCTCCGGGCATCCGCTT 6160
6161 ACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTGACAGAGTTTTTACCCGTCATCCCGAAACGCGCGAGGAGCT 6240
6241 CGGTAAGCTCATCAGCGTGGTCTGAAGCGATTTCACAGATGCTGCCTGTTTTCATCCGGTCCAGCTCGTTGAGTTTCTC 6320
6321 CAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTCTCCTGTTTGGTCACTGATGCCTCCG 6400
6401 TGTAAAGGGGATTTCTGTTCATGGGGTAAATGATACCGATGAAACGAGAGAGGATGCTACGATACGGGTTACTGATGAT 6480
6481 GAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACCACTGGCGGTTATGGATGCGGCGGGACAGAGAAAAATCACTCA 6560
6561 GGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGTTAGCCAGCAGTCCCTGCATCGATCGATCCGGA 6640
6641 ACATAATGGTGCAGGGCGCTGACTTCCGCGTTTTCCAGACTTTTACGAAACACGGAAACCGAAGACCATTTATGTTGTTGCT 6720
6721 CAGGTCCGAGACGTTTTGCGAGCAGCTCGCTTACGTTCCGCTGCGTATCCGGTATTCTGCTAACCGAGTAAGGCA 6800
6801 ACCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCAGATCATGCGCACCCGTTGGCCAGGACCCCAACGCTGCCCGAGA 6880
6881 TCGCCCGGTCGGCTGTGGAGATGGGGGACGAGATGTTTCTGCCAAGGTTGGTTTGGCATTTCACAGTTT 6960
6961 CGCAAGAATTGATTGGCTCCAATTTCTTGGAGTGGTGAATCCGTTATGCGAGGTGCGCCGGCTTCCATTAGGTCGAGGTG 7040
7041 GCCCGGCTCCATGCACCCGACGCAACGCGGGGAGGCAGACAAGGTATAGGGCGGCGCTACAATCCATGCCAACCCGTT 7120
7121 CCATGTGCTCGCCGAGGCGGCATAAATCGCCGTGACGATCAGCGGTTCCAGTATCGAAGTTAGGCTGGTAAAGACCGCGCA 7200
7201 GCGATTTGAAGCTGTCCCTGATGGTCTCATCTACCCTGCCGTCGACAGCATGGCCGTCACACCGGGCATCCGATGCCG 7280
7281 CCGGAAGCGGAGAAGAAATCATAATGGGGAAGGCCATCCAGCCTCCGCTCGGGCCCATGCCCGGCATTAATGGCCTGCTTC 7360
7361 TCGCCGAAACGTTTGGTGGCGGACCAAGTACGAAAGGCTTGGCGAGGGCGTGAAGATTCGGAATACCGCAAGCGACAG 7440
7441 GCCGATCATCGTCCGCTCCAGCGAAAGCGGTCCTCGCCGAAATGACCCAGAGCGCTGCCGGCACCTGTCTACGAGATT 7520
7521 GCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTTCATGCCCCGCGCCACCGGAAGGAGCTGACTGGGTTGAAG 7600
7601 GCTCTCAAGGGCATCGGTCCGAGATCCCGGTGCCAATAGTGTAGCTAACTTACATTAATGTCGTTGCGCTCACTGCCCGC 7680
7681 TTTCCAGTCCGGAAACCTGTCTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTTCGCTATTGGGC 7760
7761 GCCAGGTTGGTTTTTCTTTTCCAGTGTGACGGGCAACAGCTGATTTGCCCTTCCAGCCCTGGCCCTGAGAGAGTTGCGA 7840
7841 CAAGCGGTTCCAGCTGGTTTGCCTCAGCAGGCAAAATCCTGTTTGTAGTGGTGAACGCGGGATATAACATGAGCTGT 7920
7921 CTTCCGTTATCGTTCGATCCCACTACCGAGATATCCGCAACCGCGACGCGCCGACTCCGTAATGCGGCTCACTGCCCGC 8000
8001 AGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTACGATTTGCTGTTTGTGAAAACC 8080
8081 GGACATGGCACTCCAGTCCGCTTCCCGTTCGCTATCGGCTGAATTTGATTGCGAGTGTAGATATTTATGCCAGCCAGCCA 8160
8161 GACGCAGCGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCAATGCGACAGATGCTCC 8240
8241 AGCCCCAGTCCGCTACCGTCTTTCATGGGAGAAATAATAGTGTAGCTGTTGATGGTGTCTGGTGTCTGGTCAAGAAATAACGC 8320
8321 CGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCTGGTTCATCCAGCGGATAGTTAATGATCAGCCACTGACGC 8400
8401 GTTGCGCGAGAAGATTGTGCACCGCGCTTTACAGGCTTCGACGCCGCTCGTTTCTACCATCGACACCACCAGCTGGCA 8480
8481 CCCAGTTGATCGGCGGAGATTTAATCGCCGCAAAATTTGCGACGGCGGTCGAGGGCCAGACTGGAGTTGGCAACGCC 8560
8561 AATCAGACAACGACTGTTTGCCTCCGAGTGTGTTGTCGCCAGCGGTTGGGAATGTAATTCAGCTCCGCGCTCCGCTTCCA 8640
8641 CTTTTTCCCGCTTTTTCGAGAAACGTTGGCTGGCTGGTTACCACGCGGAAACGGTCTGATAAAGACACCGGCATAC 8720
8721 TCTGCGACATCGTATAACGTTACTGGTTTACATTCACCACCTGAAATGACTCTCTTCCGGGCGCTATCATGCCATACC 8800
8801 GCGAAAGGTTTTGCGCCATTGATGGTGTCCGGATCTCGACGCTCTCCCTTATGCGACTCTGCAATTAGGAAGCAGCCC 8880
8881 AGTAGTAGGTTGAGGCCGTTGAGCACCCCGCCGCAAGGAATGTT 8925

6. IVTS22

M13-Fwd T-Site 22 siteA-141 (F) M13-Rev ColEI origin AmpR

1 ACTGCCGGCCCTCTTGCGGGATATCGTCCATTCCGACAGCATCGCCAGTCACTATGGCGTGTCTGCTAGGCCATTTCGCCA 80
81 TTCAAGGCTACGCAACTGTTGGGAAGGGGATCGGTGCGGGCCTTTCGCTATTACGCCAGCTGGGGAAGGGGGATGTGC 160
161 TGCAAGCGGATTAAGTTGGGTAACGCCAGGTTTTTCCAGTACGACGTT**TGTAACACGACGGCCAGT**GAATTGCCGGCGA 240
241 TATCGGATCCATATGACGTGACGCGCTGCGAAGCTTCTAGAATTC**AGATGTGGAAACGGAAGAGCTATAATTTATAA** 320
321 **TATTTGCTTCTCCGTTTCCACATCT**GAGCTCCCGGGTACCATGGCATGCATCGATAGATCCGTCGACCTGCAGGGGGGG 400
401 GGGGAAAGCCACGTTGTGTCTCAAAATCTGTGATGTTACAT**GCACAAGATAAAAATATATCATGA**ACAATAAAAC 480
481 TGTCTGCTTACATAAAACAGTAATACAAGGGGTGTTATGAGCCATATTTCAACGGGAAACGCTTTGCTCGAGGCCGATT 560
561 AATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGGATAATGTGGGCAATCAGGTGCGACAATCTATCG 640
641 ATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAGGTAGCGTTGCCAATGATGTTACAGATGAGA 720
721 TGGTCAGACTAACTGGCTGACGAATTTATGCTCTTCCGACCATCAAGCATTTTATCCGTACTCCTGATGATGCATGG 800
801 TTACTCACCCTGCGATCCCGGGAAACAGCATTCCAGGATTTAGAAGAAATCCTGATTCAGGTGAAATAATTTGTTGA 880
881 TGGCTGCGAGTGTCTTGCCTGCGCGGTTGCATTCGATTCCTGTTTGAATGTCTTTTAAACAGCGATCGCGTATTTCTG 960
961 TCGCTCAGGCGCAATCAGCAATGAATAACGGTTTTGGTTGATGCGAGTATTTTGTGACGAGCGTAATGGCTGGCCTGTT 1040
1041 GAACAAGTCTGGAAAGAAATGCATAAGCTTTTGCATCTCACCAGGATTCAGTCTCACTCATGGTGAATTTCTCACTTGA 1120
1121 TAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGATTTGATGTTGGACGAGTCCGAAATCGCAGACCATCAAGGATC 1200
1201 TTGCCATCCTATGGAATGCCTCGGTGAGTTTTTCTCCTTCAATACAGAAACGGCTTTTTCAAATAATGGTATTGATAAT 1280
1281 CCTGATATGAATAAATTCAGATTTTCAATGATGCTCGATGAGTTTTTCTAATCAGAATGGTTAATGGTTGTACACTG 1360
1361 GCAGAGCATTACGCTGACTTGACGGGACGGCGGCTTTGTTGAATAAATCGAATTTTGTGAGTTGAAGGATCAGATCAC 1440
1441 GCATCTTCCGACAACGACAGCCGTTCCGTTGGCAAAGCAAAAGTTCAAAGATCACCACCTCCACCTCAAAATAAAGTTC 1520
1521 TCATCAACCGTGGCTCCCTCACTTCTGGCTGGATGATGGGGCATTCAGGCCGGTATGAGTACGACAACCTTCTTCA 1600
1601 CGAGGCGAGACCTCAGCGCCCCCCCCCTGCAGGTCAGGATCCATATGACGTCGACGCGTCTGCAGAAGCTTCTAGA 1680
1681 ATTC**AGATGTGGAAACGGAAGAGCTATAATTTATAAATTTGCTTCTTCCGTTTCCACATCT**GAGCTCCCGGGTACCATG 1760
1761 GCATCGATCGATAGATCTCGAGGCCCTCGCGAGCTTGGCGTAA**TCATGGTCAATAGCTGTTCTCTGTGTAATTTGTTATCC** 1840
1841 **GCTCA**CAATTCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAAGTAACTCACAT 1920
1921 TAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCCGGAAACCTGCTGTCAGCTGCATTAATGAATCCGCCAACCGCGG 2000
2001 GGGAGAGGCGGTTTTCGATTTGGCGCTTCTCCGCTTCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCCGGTCCGGCGA 2080
2081 CCGGTATCAGCTCACTCAAAGCGGTAATACGGTTATCCACAGAATCAGGGGATAACCGAGAAAGAACATGTGAGCAAA 2160

2161 AGGCCAGCAAAAAGCCAGGAACCGTAAAAAAGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCTGACGAGCATC 2240
 2241 ACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAGCTCC 2320
 2321 CTGCTGCGCTCTCCTGTCCGACCCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTCCGGGAAGCGTGGCGCTTTC 2400
 2401 TCAATGCTCACGCTGTAGGTATCTCAGTTCGGGTAGGTTCGTTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCGCTTC 2480
 2481 AGCCCCACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCA 2560
 2561 GCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGTACAGAGTCTTGAAGTGGTGGCCTAACTACGGCTA 2640
 2641 CACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCG 2720
 2721 GCAACAACCCCGCTGGTAGCGGTGGTTTTTTTGTTCGCAAGCAGCAGATTACGCGCAGAAAAAAGGATCTCAAGAA 2800
 2801 GATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATC 2880
 2881 AAAAAGGATCTTACCTAGATCCTTTTAAATTAATAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGT 2960
 2961 CTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTTCGTTTCATCCATAGTTGCCTGACTC 3040
 3041 CCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGTGCAATGATACCCGCGAGACCCAGCCTC 3120
 3121 ACCGGCTCCAGATTTATCAGCAATAAACCCAGCCAGCCGGAAGGGCCGAGCGCAGAAAGTGGTCTGCAACTTTATCCGCCT 3200
 3201 CCATCCAGTCTATTAATTGTTGCGGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATT 3280
 3281 GCTACAGGCATCGTGGTGTACGCTCGTCTGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGGCAGTTAC 3360
 3361 ATGATCCCCCATGTTGTGCAAAAAGCGGTTAGCTCCTTCGGTCTCCGATCGTTGTGCAAGTAAGTTGGCCGAGTGT 3440
 3441 TATCACTCATGGTTATGGCAGCACTGCATAATTCTTCTACTGTGCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGA 3520
 3521 TACTCAACCAAGTCATTTGAGAATAGTGTATGCGGGCACCAGTGTGCTTTGCCCGGCGTCAATACGGGATAAATACCGC 3600
 3601 GCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACCTCTCAAGGATCTTACCGCTGT 3680
 3681 TGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTACTTTACCAGCGTTTCTGGGTGA 3760
 3761 GCAAAAACAGGAAGGCAAAATGCCGCAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCTTTTT 3840
 3841 TCAATATTATTGAAGCATTTATCAGGGTATTGTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAA 3920
 3921 TAGGGGTTCCGCGCACATTTCCCGAAAAGTGCCACCTG 3959